Mathematical Modeling of T-Cell Experimental Data

THÈSE N^O 4881 (2010)

PRÉSENTÉE LE 19 NOVEMBRE 2010

À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS
LABORATOIRE POUR LES COMMUNICATIONS INFORMATIQUES ET LEURS APPLICATIONS 2
PROGRAMME DOCTORAL EN INFORMATIQUE, COMMUNICATIONS ET INFORMATION

ÉCOLE POLYTECHNIQUE FÉDERALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Irina BALTCHEVA

acceptée sur proposition du jury:

Prof. M. Hasler, président du jury
Prof. J.-Y. Le Boudec, directeur de thèse
Prof. R. J. de Boer, rapporteur
Dr C. Kohl, rapporteur
Prof. J.-P. Kraehenbuehl, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2010

Abstract

T lymphocytes (T cells) are key components of the adaptive immune system. These cells are able to recognize an enormous variety of pathogens thanks to the great specificity of their trans-membrane proteins, the T cell receptors (TCRs). TCR diversity is created during T cell maturation in the thymus by somatic gene-segment rearrangements and random nucleotide additions or deletions. Out of all possible T cell clones bearing specific TCRs, only a small fraction are successfully released in peripheral blood as the result of clonal selection. Among the selected clones, some self-reactive cells with the capacity to induce an auto-immune disease are erroneously released in periphery. To compensate for this functional flaw, the immune system has developed peripheral control mechanisms. One of them are regulatory T cells that are specialized in the control of harmful self-reactive clones. In this thesis, we combine mathematical modeling and experimental data to address immunological questions related to the dynamics of regulatory T cells and to the measurement of the structural diversity of T cell receptors. The dissertation is split into two main parts.

In the first part, we model the lifelong dynamics of human regulatory T cells ($\mathrm{T_{regs}}$). Despite their limited proliferation capacity, $\mathrm{T_{regs}}$ constitute a population maintained over the entire lifetime of an individual. The means by which $\mathrm{T_{regs}}$ sustain a stable pool *in vivo* are controversial. We define a novel mathematical model that we use to evaluate several biological scenarios about the origins and the proliferation capacity of two subsets of $\mathrm{T_{regs}}$: precursor CD4^{+}CD25^{+}-CD45RO^{-} and mature CD4^{+}CD25^{+}CD45RO^{+} cells. The lifelong dynamics of $\mathrm{T_{regs}}$ are described by a set of ordinary differential equations, driven by a stochastic process representing the major immune reactions involving these cells. Most of the parameters are considered as random variables having an *a priori* distribution. The likelihood of a scenario is estimated using Monte Carlo simulations. The model dynamics are validated with data from human donors of different ages. Analysis of the data led to the identification of two properties of the dynamics: (a) the equilibrium in the CD4^{+}CD25^{+} $\mathrm{T_{regs}}$ population is maintained over both precursor and mature $\mathrm{T_{regs}}$ pools together, and (b) the ratio between precursor and mature $\mathrm{T_{regs}}$ is inverted in the early years of adulthood. Then, using the model, we identified four biologically relevant scenarios that have the above properties: (1) if the unique source of mature $\mathrm{T_{regs}}$ is the antigen-driven differentiation of precursors that acquire the mature profile in the periphery, then the proliferation of $\mathrm{T_{regs}}$ *is essential* for the development and the maintenance of the pool; if there exist other sources of mature $\mathrm{T_{regs}}$, such as (2) a homeostatic regulation, (3) a thymic migration, or (4) a peripheral conversion of effectors into $\mathrm{T_{regs}}$, then the antigen-induced proliferation is *not* necessary for the development of a stable pool of $\mathrm{T_{regs}}$.

In the second part of the dissertation, we address the general question of TCR diversity by improving the interpretation of AmpliCot, an experimental technique that aims at the diversity measurement of nucleic acid sequences. This procedure has the advantage over other cloning and sequencing techniques of being time- and expense- effective. In short, a fluorescent dye that binds double-stranded DNA is added to a sample of PCR-amplified DNA. The sample is melted, such that the DNA becomes single-stranded, and then re-annealed under stringent conditions. The annealing kinetics, measured in terms of fluorescence intensity, are a function of the diversity and of the concentration of the sample and have been interpreted assuming second order kinetics. Using mathematical modeling, we show that a more detailed model, involving heteroduplex- and transient-duplex formation, leads to significantly better fits of experimental data. Moreover, the new model accounts for the diversity-dependent fluorescence loss that is typically observed. As a consequence, we show that the original method for interpreting the results of AmpliCot experiments should be applied with caution. We suggest alternative methods for diversity extrapolation of a sample.

Keywords: Mathematical modeling, regulatory T cells (T_{regs}), T cell receptor (TCR) diversity, AmpliCot, ordinary differential equations, likelihood estimation.

Résumé

Les lymphocytes T (cellules T) sont des éléments clés du système immunitaire adaptatif. Ces cellules reconnaissent une large variété de pathogènes grâce à la spécificité de leurs protéines trans-membranaires, les récepteurs des cellules T (TCR). La diversité des TCR est créée lors de la maturation des cellules T dans le thymus par des réarrangements de segments de gènes et l'ajout ou la suppression aléatoire de nucléotides. Parmi tous les clonotypes possiblement générés ainsi, seule une petite fraction est sélectionnée pour sortir du thymus et rejoindre les tissus périphériques en tant que population fonctionnelle. Cependant, parmi les clones sélectionnés, certaines cellules auto-réactives, ayant la capacité d'induire une maladie auto-immune, sont libérées dans la périphérie. Pour compenser cette faille fonctionnelle, le système immunitaire a développé des mécanismes de contrôle périphérique dont font partie les lymphocytes T régulateurs. Ces cellules sont en effet spécialisées dans le contrôle des clones autoréactifs possiblement nuisibles. Cette thèse combine la modélisation mathématique à des données expérimentales pour répondre à des questions immunologiques concernant la dynamique des cellules T régulatrices et l'estimation de la diversité structurelle des TCR. La dissertation consiste en deux parties principales.

Premièrement, nous étudions la dynamique des cellules T régulatrices (T_{regs}). En dépit de leur capacité de prolifération limitée, ces dernières constituent une population qui persiste pendant toute la vie d'un organisme humain. Les mécanismes par lesquels les T_{regs} se renouvellent et maintiennent une réserve stable *in vivo* sont controversés. Nous définissons un modèle mathématique afin d'évaluer plusieurs scénarios biologiques concernant les origines et la capacité de prolifération de deux sous-ensembles de T_{regs} : les cellules T_{regs} précurseurs ($CD4^+CD25^+$-$CD45RO^-$) et matures ($CD4^+CD25^+CD45RO^+$). La dynamique des T_{regs} est décrite par des équations différentielles ordinaires couplées à un processus stochastique qui simule les réactions immunitaires majeures impliquant ces cellules. La majorité des paramètres est considérée aléatoire, ayant une distribution déterminée à priori (approche du type Bayesien). Le rapport de vraisemblance entre deux scénarios est estimé par des simulations Monte Carlo. Les trajectoires du modèle sont validées par des données expérimentales provenant de donneurs humains sains de différents âges. L'analyse des données a permis l'identification de deux propriétés : (a) l'équilibre dans la population des T_{regs} opère sur les deux sous-populations de cellules précurseurs et matures prises ensemble, et (b) le rapport entre les précurseurs et les T_{regs} matures est inversé au cours des premières années de l'âge adulte. À l'aide de notre modèle, nous avons identifié quatre scénarios biologiquement réalistes qui possèdent les propriétés précédentes : (1) si la seule source de T_{regs} matures est la différentiation engendrée par un stimulus antigénique des précurseurs dans les tissus périphériques, alors la prolifération des T_{regs} *est essentielle* au développement et au maintien de leur population ; s'il existe d'autres sources de $\mathrm{T}_{\mathrm{regs}}$ matures, telles que (2) la prolifération homéostatique, (3) la migration du thymus, ou (4) la transformation de cellules effectrices en $\mathrm{T}_{\mathrm{regs}}$, alors la prolifération en réponse à une stimulation antigénique *n'est pas nécessaire* au développement d'un bassin stable de $\mathrm{T}_{\mathrm{regs}}$.

Dans la deuxième partie de la thèse, nous abordons la question générale de la diversité des TCR en améliorant l'interprétation d'AmpliCot, une technique expérimentale destinée à la mesure de la diversité des séquences d'acides nucléiques. Cette procédure présente l'avantage par rapport aux autres techniques de clonage et de séquençage d'être plus efficace en termes de coût et de temps. En bref, un colorant fluorescent, qui lie l'ADN double-brin, est ajouté à un échantillon d'ADN amplifié par PCR. L'échantillon est fondu, de sorte que l'ADN devient simple-brin, puis re-associé dans des conditions expérimentales strictes. La cinétique de ré-association, mesurée en termes d'intensité de fluorescence, est une fonction de la diversité et de la concentration de l'échantillon. Elle a été interprétée en assumant une cinétique du second ordre. À l'aide de modélisation mathématique, nous montrons qu'un modèle plus détaillé, comportant des hétéroduplexes et des complexes transitoires, reproduit de façon plus fidèle les données expérimentales. En outre, le nouveau modèle permet d'expliquer une perte de fluorescence généralement observée et qui est positivement corrélée à la diversité. En conséquence, nous montrons que la méthode d'interprétation des résultats expérimentaux, suggérée initialement, devrait être appliquée avec prudence. Nous suggérons des méthodes alternatives d'extrapolation de la diversité d'un échantillon.

Mots clés : Modélisation mathématique, lymphocytes T régulateurs ($\mathrm{T}_{\mathrm{regs}}$), diversité des récepteurs des cellules T (TCR), AmpliCot, équations différentielles ordinaires, estimateur de vraisemblance.

Contents

List of Abbreviations

APC	Antigen presenting cell
cDNA	complementary DNA
CDR3	Complementarity determining region 3
CI	Confidence interval
CM	Complete model
Cot	Concentration × time
DNA	Deoxyribonucleic acid
dsDNA	Double-stranded DNA
MLE	Maximum likelihood estimation
ODE	Ordinary differential equations
PDE	Partial differential equations
QSS	Quasi-steady state
SOK	Second order kinetics
ssDNA	Single-stranded DNA
T_{regs}	Regulatory T cells
TCR	T cell receptor

Chapter 1

General Introduction

In this thesis, we address two immunological questions by using mathematical modeling: the lifelong dynamics of regulatory T cells and the diversity measurement of a T cell receptor sample. Both questions are treated independently, but they are linked on general immunological grounds. This introduction provides the necessary immunology background and bridges the gap between both topics. A detailed introduction related to each question can be found in the corresponding chapter.

1.1 Adaptive Immunity: from Diversity to Self-Tolerance

The immune system is composed of several sophisticated and efficient defense mechanisms that have evolved to protect living organisms against pathogen attacks. The adaptive immune system is specific to vertebrates and has the ability, thanks to very specialized cells, to recognize and remember a particular invader, while it is able to tolerate “self".

One of the main actors of adaptive immunity are **T lymphocytes**. Also called T- or white blood cells, these cells exist in various types and perform different tasks. For example, cytotoxic (or killer) T cells eliminate virus-infected cells by inducing cell death of the target; helper T cells are responsible for orchestrating an immune response by secreting soluble factors, cytokines, which help and control the activation of killer T cells.

A key feature of adaptive immunity is its **diversity** that renders possible an appropriate response to practically any pathogen. One way to achieve this adaptability is through the **T cell receptor (TCR)**, a cell surface protein that allows T cells to recognize specific antigens. The TCR-mediated recognition requires that antigens are processed and presented in the form of peptides on self-MHC molecules of antigen presenting cells (APCs). In order to be able to mount an immune response to any pathogen, our bodies contain a large number of T cells with distinct TCRs. Immature T cells are produced in the bone marrow and their development continues in the thymus (hence the name “T cells"). There, immature T cells create their unique TCRs that need to pass through a series of “security checks" before cells are released in the peripheral blood as functional naive lymphocytes. The “security checks" are named *clonal selection* (Burnet, 1959) and are crucial for the good functioning of the immune system, as one of their goals is to delete the self-reactive cells that may induce auto-immune diseases.

1.1.1 Generation of TCR diversity

The T-cell receptor is a heterodimer, i.e., a protein composed of two different chains. According to the type of these chains, T cells are divided in two families: $\alpha\beta$ T cells are those that bear the α- and β- TCR chains, and $\gamma\delta$ T cells, those composed of γ- and δ- chains. $\alpha\beta$ T cells are further divided in two subsets based on their expression of the CD4 or CD8 surface molecules (or co-receptors). CD8$^+$ T cells are the cytotoxic T cells. They recognize peptides presented on MHC class I molecules and kill target cells bearing antigens recognized by their TCR through intercellular contact. CD4$^+$ T cells are the helper T cells. They recognize peptides presented on MHC class II molecules and respond to antigen by releasing cytokines that in turn help other immune cells to combat pathogens.

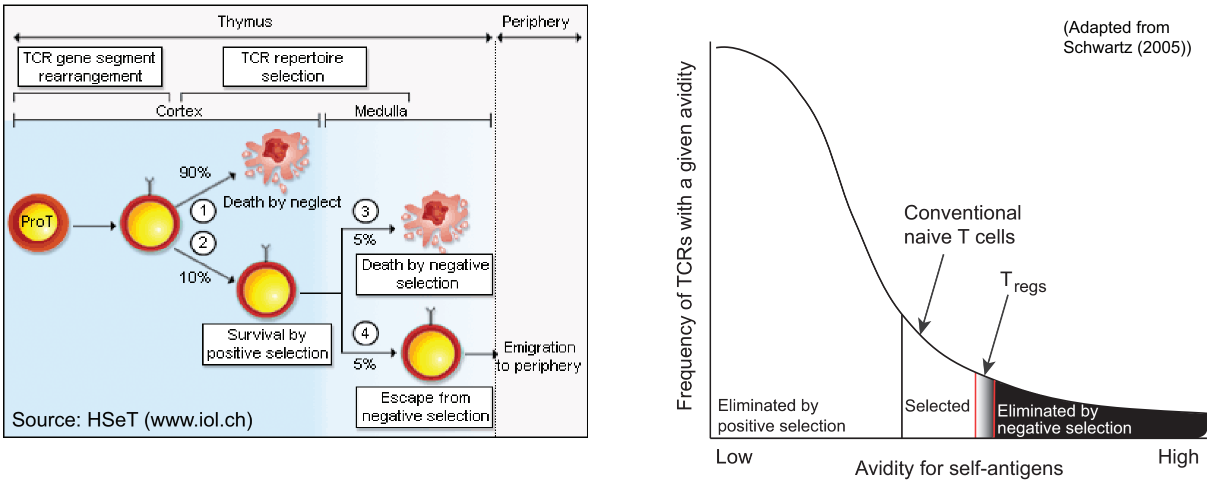
Both α- and β- chains of the TCR are partitioned into a constant (Cα or Cβ) and a variable (Vα or Vβ) regions. The constant regions are anchored in the cell membrane, whereas the variable regions are those that enter in contact with the MHC-peptide complex. Each TCR chain is encoded by multiple gene segments and specificity is created by somatic (random) rearrangement of these segments. The result of this rearrangement is a DNA sequence unique to a cell and its progeny. A large number of configurations, or clonotypes, is possible and a small number of each clonotype is released in the peripheral blood from the thymus. The notion of **diversity** used throughout this thesis refers to the number of distinct TCR clonotypes that are part of the peripheral blood repertoire. This is also stated as the structural diversity.

The theoretical number of TCR gene combinations is about 10^{18} in humans (Janeway et al., 2005). However, only a small subset of all these possible rearrangements are effectively found in peripheral blood: about 2.5×10^7 $\alpha\beta$ TCRs in humans (Arstila et al., 1999). The reasons for the large disparity between the theoretical and the effective numbers are multiple. First, there are about 10^{12} T cells in an adult male (Clark et al., 1999), thus there would be a space constraint if all successful TCR rearrangements were exported from the thymus. Second, clonal selection eliminates a large proportion of the immature thymocytes; only about 3% of all thymocytes are released in the periphery (Sompayrac, 2003). Third, the above diversity estimate (2.5×10^7) is only a lower bound; currently, the diversity of human T cells is unknown. However, it has been shown that a loss of repertoire diversity (with respect to a healthy state) is associated with disease or aging. Studying the repertoire diversity is hence important for understanding pathological states.

Several experimental techniques aim at the measurement of the structural TCR diversity of a repertoire sample. For example, Immunoscope (or Spectratype) gives a qualitative insight of the repertoire's shape in terms clonal sizes (Currier and Robinson, 2001; Pannetier et al., 1993); high-throughput DNA sequencing exhaustively enumerates the clonotypes of a sample, thus providing a more detailed picture of the repertoire (Mardis, 2008; Shendure and Ji, 2008). AmpliCot is an alternative experimental technique that allows the sample diversity measurement through quantitation of the re-hybridization speed of denatured PCR products (Baum and McCune, 2006). This elegant approach has the advantage over the cloning and sequencing methods to be time- and expense- effective. However, in order to obtain accurate diversity estimates, the assay should be performed under very stringent experimental conditions.

1.1.2 Clonal Selection

Clonal selection occurs in two steps. The first step is *positive selection*. As T cells recognize processed peptides presented on self-MHC molecules, TCRs need to be compatible with self MHC-peptide complexes. At the stage of positive selection, cells whose TCRs do not have a sufficient binding affinity (or equivalently, avidity) with self MHC-peptide complexes are eliminated. The second step is *negative selection*. Most of the self-reactive clones, which have a too strong affinity to self MHC-peptide complexes, are eliminated at this stage. Only a small proportion of all produced T cells survive positive and negative selection (Figure 1.1A).


Thymus
Periphery
TCR gene segment rearrangement
TCR repertoire selection
Cortex
Medulla
ProT
90%
1
Death by neglect
2
10%
3
5%
Death by negative selection
Survival by positive selection
4
5%
Emigration to periphery
Escape from negative selection
Source: HSeT (www.iol.ch)
(Adapted from Schwartz (2005))
Frequency of TCRs with a given avidity
Conventional naive T cells
T_{regs}
Eliminated by positive selection
Selected
Eliminated by negative selection
Low
Avidity for self-antigens
High

A. Qualitative view

B. Quantitative view

Figure 1.1: Clonal selection: the link between TCR diversity and regulatory T cells. T cells are selected in the thymus according to their TCR. Only a small number of maturating T cells survive positive and negative selection and are released in the peripheral blood as functional naive lymphocytes. A: qualitative view of the clonal selection process; B: quantitative view revealing one possible way of generating natural T_{regs} (the “instructive" model). According to this model, T_{regs} would originate from “slightly" self-reactive thymocytes whose TCR avidity is in the grey zone on the frontier between positively and negatively selected clones. Figure source: panel A: from HSeT, www.iol.ch, panel B: adapted from Schwartz (2005).

However, clonal selection is not perfect. Intuitively, if cell fate is determined by the interactions between TCRs and self-antigens presented on MHC molecules, then the TCR repertoire is shaped according to the self-antigens that are presented in the thymus during the maturation process. In view of the large amount of self-antigens, it is easy to imagine that not all self-antigens are presented in a particular time frame, hence some self-reactive TCR clonotypes can be released in peripheral tissues. Additional tolerance mechanisms are therefore needed to control such self-reactive cells. Among those reported, we find receptor editing (McGargill et al., 2000) and functional inactivation (anergy) (Ramsdell and Fowlkes, 1990), which both occur in the thymus. Despite these central tolerance mechanisms, some self-reactive cells still escape clonal selection and find their way through to periphery. Fortunately, there are peripheral tolerance mechanisms; **regulatory T cells ($\mathbf{T_{regs}}$)** are one of these.

1.1.3 Regulatory T cells

Regulatory T cells (T_{regs}) are a subset of $CD4^+$ helper T cells. These specialized "controllers" act as antagonists to immune responses by suppressing the activation and proliferation of $CD4^+$ helper and $CD8^+$ killer T cells. By this means, T_{regs} are involved in self-tolerance (Kim et al., 2007), homeostasis and in the control of excessive immune reactions. Several types of cells exhibit regulatory functions, for example the induced T_{regs}, the IL-10 secreting T_R1 cells, the TGFβ-secreting T_H3 cells, or the adaptive T_{regs}. Here, we focus mainly on the so-called *naturally occurring T_{regs}*.

Naturally occurring T_{regs} are identified by the surface expression of CD4, as well as by the expression of the transcription factor forkhead box P3 (FoxP3), a negative modulator of IL-2 transcription (Fontenot et al., 2003; Hori et al., 2003). These cells are generated in the thymus through mechanisms that are not yet fully understood. Their development is programed and controlled by the master regulator FoxP3 and requires high-affinity interactions between their TCR and self MHC-peptide complexes (Picca et al., 2006). Similarly to the commitment to CD4+/CD8+ cell lineages, or the differentiation pathways of effector/memory cells, two models of thymic T_{regs} generation have been proposed: a deterministic “instructive" model (Jordan et al., 2001; Lio and Hsieh, 2008; Modigliani et al., 1996) and a “stochastic" selection model (van Santen et al., 2004). In the first, immature thymocytes that are potentially self-reactive commit to the regulatory lineage. The concept can be visualized by considering the distribution of TCR avidities, result of the TCR gene segment rearrangements (Figure 1.1B). The low avidity cells are eliminated by positive selection, whereas the high avidity cells are deleted by negative selection. The selected naive (non-regulatory) T cells have TCR avidity that is neither too high, nor too low. According to the instructive selection model, T_{regs} would have TCR avidity in the high end of the avidity spectrum because they originate from potentially self-reactive clones (Schwartz, 2003). In the stochastic selection model, the T_{regs} lineage conversion signal would be rather stochastic, independent of TCR avidity.

T_{regs} ontogeny therefore suggests that these cells are either created as a separate lineage through unknown factors independent of TCR specificity (stochastic selection), or they originate from a thymocyte ancestor common to conventional $CD4^+$ T cells, where TCR specificity plays a key role (instructive selection). These two distinct pathways affect the way T_{regs} are viewed in terms of proliferation capacity. Indeed, the paradigm of central tolerance in the thymus states that self-reactive clones are either deleted (Burnet, 1959), or rendered functionally inactive (anergic) (Ramsdell and Fowlkes, 1990). If the instructive selection model is accepted, it would be logical to think that T_{regs} are anergic cells, thus have a limited proliferation capacity. If, instead, the stochastic selection model is accepted, there is no reason to think that T_{regs} would have an impaired proliferation capacity. The dichotomous nature of their ontogeny is probably at the source of what we call the T_{regs} “proliferation controversy".

A major challenge in current studies of regulatory T cells is the lack of cell surface markers proper to this population. The expression of the transcription factor FoxP3, which identifies T_{regs}, is intracellular and consequently necessitates the destruction of cells for detection. Moreover, FoxP3 does not solely identify thymus-derived T_{regs}, but also cells that have acquired suppressive phenotype and functions in the periphery (Curotto de Lafaille et al., 2004;

Vukmanovic-Stejic et al., 2006) [1]. The cell surface marker mostly used hitherto is CD25, the α-chain of the IL-2 receptor, which has been shown to be constitutively expressed on T$_{\text{regs}}$ (Baecher-Allan et al., 2001; Miyara et al., 2009b; Sakaguchi et al., 1995; Taams et al., 2002). However, CD25 is also expressed transiently on activated (non-regulatory) T cells, which renders the isolation of T$_{\text{regs}}$ difficult. Several other markers have been associated with naturally occurring T$_{\text{regs}}$, notably high levels of CTLA-4 (cytotoxic T-lymphocyte associated molecule-4), CD62L, CCR7, GITR (glucocorticoid-induced TNF receptor) and low levels of CD127 (the alpha-chain of the IL-7 receptor) (Codarri et al., 2007; Liu et al., 2006; Seddiki et al., 2006a). Currently, the combination of high or low expression levels of several of the above surface proteins are used to detect T$_{\text{regs}}$ within a reasonable degree of purity (reviewed in Sakaguchi et al. (2010)). Hitherto, a typical experimental approach consists in isolating CD4$^+$CD25$^+$ T cells, performing the desired experiment, and verifying *a posteriori* the enrichment of T$_{\text{regs}}$ in the sample by immobilizing the cells.

Robust T$_{\text{reg}}$ cell identification is one among many unresolved issues related to these cells. Although important advances have been recently made, the main molecular mechanism(s) of T$_{\text{regs}}$-mediated suppression in humans remain elusive. It is established that T$_{\text{regs}}$ need to be activated through their TCR to be functionally suppressive and the strength of TCR stimulation influences the effectiveness of suppression. Moreover, suppression can be contact-dependent or cytokine-mediated (see Sakaguchi et al. (2010) for a review). Despite the remaining open questions, the broad range of clinical applications of regulatory T cells (Miyara et al., 2009a; Safinia et al., 2010; Trzonkowski et al., 2009) is rendering this research area very active and intriguing.

1.2 Mathematical Modeling in Immunology

With the modern advances of experimental techniques, a large amount of immunological data is created. The complexity of interactions between individual components and the difficulty to isolate influencing factors makes the use of mathematics challenging, but valuable. Mathematical models in immunology exist since several decades. Like any other research field where two or more disciplines are crossing, the challenges of modeling a biological phenomena are multiple. The size of a living system, the number and the complexity of interactions between individual components is such that tractable and computationally efficient models are difficult to derive. Nevertheless, mathematical models have been applied to immunology since more than 50 years (Louzoun, 2007).

In this thesis, we develop mathematical models in order to gain biological insight and improve the interpretation of T-cell-related experimental data. The dissertation is divided in two parts, each part addressing the following immunological questions: (I) the *in vivo* dynamics of human regulatory T cells; (II) the measurement of the structural diversity of a TCR sample with the AmpliCot technique.

[1] Moreover, a non-regulatory FoxP3-expressing CD4$^+$ T cell population has been recently discovered in the peripheral blood of humans (Miyara et al., 2009b).

1.3 Dissertation Outline

The dissertation is structured as follows. In Chapter 2, we define a generic model that describes the lifelong dynamics of regulatory T cells. We use our model to address the T_{regs} proliferation controversy. In that perspective, we derive from the generic model several biologically plausible scenarios about the origins and the proliferation capacity of these cells. The model scenarios are challenged against human *ex vivo* data and some of them are discarded.

In Chapter 4, we address the general question of TCR diversity by improving the interpretation of AmpliCot, an experimental technique that aims at the diversity measurement of nucleic acid sequences. We use mathematical modeling to describe AmpliCot experimental data. Once again, we evaluate two model variants by fitting them to data. Practical and methodological conclusions are then drawn.

Chapter 3 and Chapter 5 are auxiliary chapters that contain theoretical and analytical results related to the T_{regs} and the AmpliCot models respectively. We make concluding remarks in Chapter 6.

Part I

Dynamics of Regulatory T cells

Chapter 2

Lifelong Dynamics of Human CD4$^+$CD25$^+$ Regulatory T Cells

2.1 Introduction

As of today, two developmental pathways of human regulatory T cells *in vivo* have been identified (Sakaguchi, 2003): naturally occurring thymus-derived T$_{\text{regs}}$ (Fritzsching et al., 2006; Hoffmann et al., 2006; Seddiki et al., 2006b; Wing et al., 2002) and adaptive or induced T$_{\text{regs}}$, derived from non-regulatory CD4$^+$CD25$^-$FoxP3$^-$ T cells (Kretschmer et al., 2005; Vukmanovic-Stejic et al., 2006; Walker et al., 2003a). Naturally occurring T$_{\text{regs}}$ originate in the thymus and are released in the periphery with a naive phenotype (Cupedo et al., 2005; Seddiki et al., 2006b; Takahata et al., 2004; Wing et al., 2002, 2003). They are identified as CD4$^+$CD25$^+$CD45RO$^-$ T cells and will be called “precursor" T$_{\text{regs}}$ throughout this chapter [1]. Once precursors encounter their cognate antigen, they acquire a suppressive capacity and a memory phenotype (Fritzsching et al., 2006). We call these differentiated cells “mature" T$_{\text{regs}}$. Adaptive T$_{\text{regs}}$ are derived from rapidly proliferating activated-effector or memory CD4$^+$CD25$^-$ T cells that acquire the permanent expression of CD25 and the suppressive function in the periphery (Vukmanovic-Stejic et al., 2006; Walker et al., 2003a). Because they have experienced antigen, these cells have a mature profile and express the memory phenotype CD45RO. It is important to remark that there is no difference regarding the surface markers characterizing both types of mature T$_{\text{regs}}$ and consequently, one can not distinguish both T$_{\text{regs}}$ origins using fluorescent techniques. The same observation can be made about T-cell receptor excision circles (TREC): T$_{\text{regs}}$ from both origins have similar decreased TREC content (Kasow et al., 2004). Thymus-derived regulatory cells divide during clonal selection in the thymus and in the periphery, whereas activation-induced regulatory cells come from rapidly proliferating non-regulatory T cells and therefore have decreased TREC content.

In vivo studies of human T$_{\text{regs}}$ have shown that the number of precursors decreases significantly with age (Seddiki et al., 2006b; Valmori et al., 2005), whereas the number of mature T$_{\text{regs}}$ increases in elderly individuals (Gregg et al., 2005). Thymic involution, together with the fact

[1] These cells are also named *naive* T$_{\text{regs}}$ (Miyara et al., 2009b).

that CD4$^+$CD25$^+$CD45RO$^+$ mature regulatory T cells are known to be non-proliferating, introduces the question of how is developed and maintained a stable pool of T$_{\text{regs}}$ throughout life. Different hypotheses are evoked in Akbar's opinion paper (Akbar et al., 2007); the question of proliferation is central. The first hypothesis claims that precursor CD4$^+$CD25$^+$CD45RO$^-$ cells are able to proliferate (Klein et al., 2003; Walker et al., 2003b) and even though the thymic involution reduces the input of newly produced precursors with age, these cells are the main reservoir of mature T$_{\text{regs}}$. The second hypothesis suggests that both precursor and mature T$_{\text{regs}}$ are non-proliferating but although mature T$_{\text{regs}}$ are sensitive to death because of their high levels of CD95 (Fritzsching et al., 2006; Taams et al., 2001), the fact that precursors are apoptosis-resistant suffices to sustain a stable pool of mature T$_{\text{regs}}$. Finally, the third hypothesis points out the presence of an external source of mature T$_{\text{regs}}$, derived from rapidly proliferating effector CD4$^+$CD25$^-$ T cells (Taams et al., 2001; Vukmanovic-Stejic et al., 2006). Thus, there is a controversy about the mechanisms by which T$_{\text{regs}}$ regenerate throughout the lifetime of individuals. The objective of our study is to evaluate the above biological hypotheses by the means of a mathematical model and to measure their effect on the development and maintenance of a pool of human T$_{\text{regs}}$.

A major difficulty encountered in mathematical modeling of biological systems is dealing with parameter values. The more detailed the model is, the more parameters it involves and its behavior can be completely different according to the values taken by the latter. Good parameter estimates exist in some cases, but it is often difficult to build an experiment that allows for their direct measurement. In the mathematical model presented hereafter, we employ a modeling technique that alleviates the parameter estimation problem by considering parameters as random variables having *a priori* distributions (as in Bayesian approaches). This approach allows us to evaluate simultaneously several values, to produce results that depend little on the exact values, and therefore to diminish the probability of errors due to wrong parameter estimates.

Using the above technique, we define a generic model describing the lifelong dynamics of T$_{\text{regs}}$. We attempt to include in it all the actual knowledge about the population dynamics of regulatory T cells, while keeping the model as simple as possible. We consider all events that affect the population size of T$_{\text{regs}}$: immune reactions to self or foreign antigens, the homeostatic activity and the external input – from the thymus or from the non-regulatory effector CD4$^+$CD25$^-$ T cell pool. We then implement the above-mentioned hypotheses about the origins and proliferation capacity of T$_{\text{regs}}$. In order to validate the model, we compare it to human data consisting in measurements of T$_{\text{regs}}$ as a function of age. We first study the expected model behavior where the stochastic parameters take the average value of their *a priori* distributions. Then, we evaluate the performance of the model with random parameters. Its behavior is evaluated for several values inside the given parameter range and the density of trajectories is estimated. This density leads to the definition of the likelihood of a model scenario, used to formally reject the scenarios that are unable to fit the data.

2.1.1 Mathematical Models of T$_{\text{regs}}$: State of the Art

As of today, a certain amount of mathematical models of regulatory T cells can be found in the literature. Most of them analyze the suppression mechanisms of these cells and their interactions with other cell types and cytokines. In this section we give an overview of the currently existing work. For each cited study, in addition to the main findings, the methodological approach is highlighted. We also mention whether $\mathrm{T_{regs}}$ were considered as proliferative or not.

Crossregulation Model of Immunity ($\mathrm{T_{regs}}$ and APCs)

The most extensively studied and developed model of regulatory T cells is certainly the Crossregulation model of immunity (León et al., 2000). In their original paper, León et al. (2000) test several suppression mechanisms by considering populations of antigen presenting cells (APCs), regulatory and conventional T cells. The authors postulate that the cell-mediated suppression occurs through the formation of multicellular conjugates of T cells and APCs. Three suppression scenarios are examined: (1) competition between $\mathrm{T_{regs}}$ and conventional T cells for conjugation sites on APCs; (2) proliferation inhibition of conventional T celly by $\mathrm{T_{regs}}$ on conjugates with APCs; (3) in addition to the proliferation inhibition, the growth of $\mathrm{T_{regs}}$ is hypothesized to be dependent on conventional T cells. The model is described by ordinary differential equations (ODE) and a quasi-steady state assumption is applied to conjugates (which significantly simplifies the equations). The authors identify different parameter regimes, according to which the equilibrium points might change (and bi-stability appears). Using a phase-plane and bifurcation analysis, each possible scenario is related to existing experimental observations (qualitatively). To account for the outcome of adoptive transfer experiments, the authors conclude that the correct suppression model should exhibit bi-stability, leading to a tolerant and an auto-immune steady state, where either $\mathrm{T_{regs}}$ or effectors dominate. The third suppression model is retained as most plausible: active growth suppression of effectors by $\mathrm{T_{regs}}$ and maintenance of $\mathrm{T_{regs}}$ dependent on effectors (through IL-2).

The same model is then applied to the analysis of experimental data of linked suppression *in vitro* (León et al., 2001). The experimental set up consists in APCs and target conventional T cells that are co-cultured with or without $\mathrm{T_{regs}}$. Target cells are stimulated and the population expansion is measured. It is hypothesized that when conjugated with $\mathrm{T_{regs}}$ simultaneously, the proliferation of conventional T cells is reduced. The inhibition index is defined as the ratio between the cell counts in a culture containing $\mathrm{T_{regs}}$ and those in a culture without $\mathrm{T_{regs}}$. In this model, $\mathrm{T_{regs}}$ are assumed to be non-proliferative and two mechanisms of suppression are examined: simple competition for APC conjugation sites versus competition + active inhibition. The authors observe that the inhibition index is mainly determined by the number of regulatory cells per APC and is insensitive to the number of target (suppressed) cells. However, they fail to fit the model to the experimental data and put forward several explanations, among which the possibility that $\mathrm{T_{regs}}$ proliferate in the presence of IL-2-producing conventional T cells.

In the subsequent study, León et al. (2003) further develop the Crossregulation model to add thymic input and the simultaneous peripheral dynamics of several T cells clones. This is done by simulating clonal selection. New clones are generated stochastically and the equations describing their dynamics are appended to the existing ODE system. If the population size of a particular clone vanishes, its corresponding equation is removed. The system is then perturbed in order to simulate two types of events: the introduction of a foreign antigen and its clearance. After each external perturbation, the new steady state is computed. This is different than our system, where immune reactions are internally generated by the stochastic process. The authors test the effect of thymectomy and observe that diversity is lost by competitive exclusion.

The Crossregulation model is used by León et al. (2004) to study the correlation between the incidences of autoimmune diseases and infection. An inverse correlation is revealed: the risk of autoimmunity is the price that must be paid for assuring immune responses to pathogens.

Carneiro et al. (2005) further study an alternative self-tolerance mechanism, mediated by the tuning of activation thresholds, which would make auto-reactive T cells reversibly "anergic" and unable to proliferate. It turns out that this hypothesis is only partially compatible with the qualitative observation of adoptive transfer experiments and was therefore left out.

The Crossregulation model was also applied to study tumor immunobiology. León et al. (2007a) provide an explanation to the observation that the development of some tumors expand regulatory T cells, whereas others do not. In a subsequent study (León et al., 2007b), the same authors consider how these two tumor classes respond to different therapies, namely vaccination, immune suppression, surgery, and their different combinations. Model responses to different therapies are simulated as particular dynamical perturbations to the ODE system.

In a review paper, Carneiro et al. (2007) gather all the above facts and observations revealed by the Crossregulation model, both theoretical and experimental, in an unifying theory in which "the persistence and expansion of T$_{\text{regs}}$ depend strictly on specific interactions they make with APCs and conventional effector T cells." Although the importance of APCs in the function of T$_{\text{regs}}$ is largely emphasized in the Crossregulation model, throughout all the papers, the actual dynamics of these cells are either considered as a parameter or as an externally controlled variable. The following work brings more attention to the explicit dynamics of APCs.

T$_{\text{regs}}$ and APCs

Alexander and Wahl (2010) model antigen-specific natural T$_{\text{regs}}$, together with a particular type of self-antigen, its corresponding APCs and the responding effectors. The authors focus on a positive feedback loop between effectors that release antigens, which are taken up by APCs that, in turn, stimulate more effector T cells. The production of T$_{\text{regs}}$ is proportional to both effectors (IL-2 producing cells) and APCs and they act by suppressing the action of APCs. Different suppression mechanisms are analyzed. Interestingly, the authors present both a deterministic and a stochastic version of the same system. The ODE system reveals bi-stability, as in most T$_{\text{regs}}$ models, with a trivial (tolerance) and non-trivial (auto-immune) stable state. The limit behavior determined by the basic reproductive ratio R_0. The stochastic version of the ODEs is derived using an approach similar to the one of Chao et al. (2004), in which the ODEs are discretized and the populations are updated at each time step by generating Binomial or Poisson random variables. Once again, the model is not challenged against experimental data. One of the conclusions is that self-antigen-specific T$_{\text{regs}}$ play no role in the system's qualitative long-term behavior, but have quantitative effects that could potentially reduce and clear an auto-immune response. The important role of T$_{\text{regs}}$ with arbitrary specificity is highlighted. Finally, the probability of developing a chronic auto-immune disease increases with the quantity of initial-exposure antigen or of auto-reactive effectors.

$\mathrm{T_{regs}}$ and Cytokine Kinetics

Burroughs et al. (2006) study regulatory and auto-immune (conventional) T cells by assuming that all cell interactions are realized through cytokines. In particular, the authors observe the consequences of $\mathrm{T_{regs}}$ inhibition of IL-2 secretion. Their model is composed of resting and activated regulatory and auto-immune T cells, as well as two cytokines: IL-2 produced by auto-immune cells and consumed by both auto-immune and regulatory T cells, and another cytokine, produced by tissues, consumed by $\mathrm{T_{regs}}$ only. $\mathrm{T_{regs}}$ inhibit the secretion of IL-2 and do not produce IL-2 themselves. It is assumed that $\mathrm{T_{regs}}$ proliferate homeostatically by competing for the tissue-secreted cytokine. In addition, these cells can also proliferate in the presence of IL-2, but less efficiently than conventional T cells. The stability analysis of the ODE system reveals a "control" state in which $\mathrm{T_{regs}}$ dominate and eliminate auto-immune cells, and an "auto-immune" state in which the latter expand and escape $\mathrm{T_{regs}}$ control. The main conclusion of this theoretical work is that the shift towards control or auto-immunity is dependent on the efficiency of auto-immune T cells to utilize IL-2 compared to $\mathrm{T_{regs}}$. This efficiency can be compensated by auto-immune cells by an increase in their number. In a later paper, Burroughs et al. (2008) further analyze the above model and provide a sensitivity analysis to the parameters.

More recenlty, Garcia-Martinez and León (2010) extend the Crossregulation model by explicitly modeling the dynamics of IL-2. By doing so, they allow for non-local/unspecific interactions between effectors and $\mathrm{T_{regs}}$ in the sense that interactions are possible not only upon simultaneous conjugation on the same APC, but also via the free IL-2 present in the same lymph node. Thus, this model combines the assumptions of the Crossregulation model and the model of Burroughs et al. (2006) and constitutes a rather complete picture of the cell interactions in a lymph node. The same methodological approach as in León et al. (2000) is applied. Two model variants, with different roles of IL-2 in suppression, are tested: (1) $\mathrm{T_{regs}}$ suppress effectors by competition for IL-2 only; (2) in addition to competition, $\mathrm{T_{regs}}$ inhibit the activation of effectors co-localized on the same APC. The authors establish parameter constraints in the extended models in order to reproduce the basic properties of the Crossregulation model (bi-stability, etc.). Furthermore, the extended models lead to new properties of the dynamics. An interesting characteristic regarding the unspecific regulation is observed. The authors consider the case of two different antigen-specific clones of effectors and $\mathrm{T_{regs}}$ responding to two sets of APCs. An abrupt increase in the number of APCs of type 1 is then applied, simulating, for example, a particular infection. In version (1) of the model, the responses to both types of APCs are fully coupled. This means that the increase in APCs of type 1 can break tolerance for the corresponding cells (effectors of type 1 take over $\mathrm{T_{regs}}$ of type 1) and the same would happen for the cells of type 2. This would lead to collateral damages if all clones are activated by the stimulation of a single one. The situation is more realistic in model variant (2), where the activation of clone of type 1 does not imply the activation of the other clone.

Adaptive/Induced $\mathrm{T_{regs}}$

Fouchet and Regoes (2008) model an interaction network composed of adaptive $\mathrm{T_{regs}}$, effector T cells and APCs. The model is very similar to one version of the Crossregulation model, except for the fact that resting APCs can induce the transformation of effectors into $\mathrm{T_{regs}}$. The authors define an ODE system in which precursor T cells may differentiate into either adaptive

T$_{regs}$ or effector T cells, depending on the activation state of the APC. An equilibrium analysis reveals the existence of two stable equilibriums (similarly to the Crossregulation model): one tolerant (regulated) state where T$_{regs}$ control effectors and one unregulated state where the vanishing population of T$_{regs}$ cannot control the effectors. Then the authors study the effect of parameters on the nature of the equilibrium regime. The bifurcation analysis reveals that the switch from the regulated to the unregulated state depends on the strength of the antigenic stimulus and the state from which the network has been perturbed.

In their experimental study of adaptive T$_{regs}$, Vukmanovic-Stejic et al. (2006) use the model developed in Macallan et al. (2003) to estimate the *in vivo* proliferation and death rates of memory-derived FoxP3+ regulatory T cells. The latter model is specifically defined for labeling experiments, where deuterium from deuterated glucose is incorporated into the DNA of dividing cells. The model has one compartment, the amount of labeled deoxyadenosine, and accounts for its appearance and disappearance due to cell proliferation and death (further deuterium labeling models can be found in Mugwagwa (2010)). The analysis of experimental data reveals the peripheral conversion or rapidly proliferating CD4$^+$CD25$^-$FoxP3$^-$ memory effectors into a regulatory CD4$^+$CD25^{hi}FoxP3$^+$ phenotype.

After developing an extremely detailed delay differential equations (DDE) model to study the role of natural T$_{regs}$ in the adaptive immune system (Kim et al., 2007), Kim et al. (2010) study the dynamics of primary T-cell responses and the possible involvement of adaptive regulatory T cells. The authors challenge the paradigm of a primary T cell response, according to which, (1) T-cell dynamics in response to an antigen do not depend on the level and duration of antigen stimulation; and (2) T-cell responses are independent of the clone size of antigen-specific responders. Using delay and partial differential equations, the authors conclude that the old "paradigm does not entirely capture the observed robustness of T cell responses to variations in precursor frequency". They propose an alternative mechanism in which the dynamics of a primary T-cell response are governed by a feedback loop involving adaptive regulatory cells rather than by intrinsic developmental programs.

T$_{regs}$ and Gene Expression

In a general setting of helper T cell differentiation, Van den Ham and de Boer (2008) propose a model describing the expression of master regulators. Master regulators are transcription factors that are both necessary and sufficient for the induction of a certain cellular phenotype. For example, to polarize a helper T cell towards the helper 1 type, Tbet is necessary, whereas for helper type 2, it is GATA3. In the case of T$_{regs}$, FoxP3 is the master regulator. The general framework of Van den Ham and de Boer (2008) has been applied to experimental data measuring the expression levels of FoxP3 and GATA3 mRNA.

T$_{regs}$ and Evolution

Saeki and Iwasa (2009) develop a mathematical model to study the advantages of having regulatory T cells in the immune system. The authors weight the pros and the cons of a robust ability to cope with foreign antigens, versus auto-immunity. Using a probabilistic approach, they define the fitness function of an organism, which takes into account the benefit of having effector T cells reactive to foreign antigens and the severity of having effectors attacking self tissues. A model of cell maturation in the thymus is presented, where immature T cells are determined to be regulatory or effectors, based on whether or not they interact with self-antigens during clonal selection (they also consider another version, where cell fate is pre-determined). It is assumed that the number of times a particular auto-reactive immature T cell meets with its corresponding antigen during clonal selection follows a Poisson distribution. Further, once in the periphery, activated T_{regs} are assumed to suppress effectors by direct interaction. The Poisson distribution is again used to model the number of times a particular effector T cell meets regulatory T cells. From this is calculated the probability of a cell not being suppressed. The fitness function of a system with T_{regs} is then compared to the fitness without T_{regs} and conclusions are drawn with respect to the parameters. Localized and global suppression are compared. It turns out that it is advantageous to have regulatory T cells if suppression is localized, i.e., "if the body is composed of many compartments, and regulatory T cells suppress the immune reactions only within the same compartment". The framework presented by the authors gives an interesting insight of T cell maturation and T_{regs} formation.

In the following paper, Saeki and Iwasa (2010) use an extension of the same mathematical model to study the optimal number of regulatory T cells. The authors propose that this number depends on the "number of self-antigens, the severity of auto-immunity, the abundance of pathogenic foreign antigens, and the spatial distribution of self-antigens in the body."

2.1.2 Our Contributions

None of the above models has studied in depth the quantitative effect of T_{regs} proliferation capacity and origins on the lifelong peripheral pool size in humans. At the time of development of our model, the proliferation capacity of T_{regs} was a controversial issue. Our first contribution is therefore to address quantitatively this question and to challenge the plausibility of several hypotheses with human *ex vivo* data.

From the perspective of proliferation capacity of T_{regs}, the model of Burroughs et al. (2006) is the closest one to ours, because it considers strict conditions for T_{regs} proliferation. However, the authors of this study did not consider the lifelong dynamics resulting from the repeated exposure to self- or foreign- antigens.

None of the above studies has sorted T_{regs} on the expression of the memory-type receptor CD45RO. Once again, the closest discrimination was in Burroughs et al. (2006), where the authors consider active and inactive T_{regs}, which was mainly translated in a suppression capacity difference. In addition to distinguishing the activated from the resting T_{regs} populations, we also account for antigen-unexperienced precursors, which were known to have only a limited suppression and a marked proliferation capacity.

A striking difference between our model and most of those studying T_{regs} dynamics is the fact that we neglect the population of APCs. However, only in Alexander and Wahl (2010) are the dynamics of APCs modeled explicitly. In all the variants of the Crossregulation model (Carneiro et al., 2007), the APCs are assumed to be in equilibrium and their presence is only expressed through one parameter, the density of APC conjugation cites. Thus, these models reduce to a predator-pray system, where either one of the conventional or regulatory T cells outcompetes the other, either both populations co-exist. Our intention is to start by building the simplest possible model that can address the relevant question. As we are not concerned with suppression mechanisms, we ignore APCs and conventional T cells. The effect of IL-2 as a growth factor is modeled implicitly through the parameters. Nevertheless, the inclusion of these components would certainly be of interest.

Our second contribution is the fact that our model is conceptually different from the existing T$_{\text{regs}}$ models. Indeed, the combination of a stochastic process with ODEs is an original approach that has not been applied to the study of T cell dynamics. In addition, we employed random parameters with a Bayesian-type statistical analysis, which is also a marginal manner of treating unknown parameter values.

Finally, we have fitted a quantitative model to time-series experimental data, which has not yet been done in the case of human regulatory T cells. Hitherto, data sets have been used to fit the time-dependent dynamics of conventional CD4+ or CD8+ T cells (De Boer et al., 2003; Mugwagwa, 2010).

2.1.3 Chapter Outline

This chapter is organized as follows. We first present in Section 2.2.1 the detailed description of the model, followed by the materials and methods necessary to the data obtention (Section 2.2.2). In Section 2.2.3, we describe how the model is fitted to the data and how the scenarios are formally evaluated. The results are in Section 4.3 and the discussion in Section 4.4.

2.2 Materials and Methods

2.2.1 $\mathbf{T_{regs}}$ Mathematical model

In order to model the biological hypotheses concerning the origins and the proliferation capacity of $\mathrm{T_{regs}}$, we consider several scenarios of a generic mathematical model. The scenarios are the result of imposing constraints on model parameters such as the input, proliferation and death rates of precursor and mature $\mathrm{T_{regs}}$ in the generic model.

We use several mathematical tools in order to describe in a robust way $\mathrm{T_{regs}}$ lifelong kinetics. Changes in the populations' sizes due to immune reactions and homeostatic activity are described by ordinary differential equations (ODEs). Events corresponding to encounters with antigen-presenting cells leading to (auto-)immune responses are generated according to a stochastic process. A sketch of the entire system with all its components can be seen in Figure 2.1. For a quick reference, Table 2.1 summarizes all the assumptions of our model. In what follows, we describe in detail the generic model, the model scenarios, the parameter fitting and the model evaluation procedure.

Generic model

The generic model describes the lifelong dynamics of $\mathrm{T_{regs}}$. We call it generic because it is taking into account all possible events that may affect $\mathrm{T_{regs}}$ population size and because it is a generalization out of which we define the model scenarios that describe the studied biological hypotheses. The generic model is composed of two parts: a deterministic part describing cell dynamics during immune responses and homeostasis, and a stochastic part generating all infection events occurring in a human lifetime.

(Auto-)Immune reactions process In order to study the dynamics of regulatory T cells over the entire lifetime of a human, responses to self- and foreign-antigens are included in our model. We consider two types of immune reactions: the minor ones, occurring frequently and triggered by self- or dietary antigens, and the major ones, occurring rather seldom, mainly provoked by foreign antigens and having a great impact on the $\mathrm{T_{regs}}$ pool. Minor immune reactions triggered by antigens sampled in the mucosa-associated lymphoid tissues are taken into account in the pool-size control dynamics described in the next section. Major immune reactions are triggered by successful interactions with antigen-presenting cells, in which a cytokine environment allowing $\mathrm{T_{regs}}$ activation and (possibly) proliferation is present (Carneiro et al., 2007). As we found no information about the frequencies and time-distribution of such immune reactions *in vivo*, we model this phenomena with a stochastic process having a constant rate. Thus, we assume the length of time-intervals between two infections to be random, having a shifted exponential distribution with mean $\lambda + \delta$, where δ is the minimal duration of an immune response. We make the simplifying assumption that during the first phase of an immune response (of length δ), no other infections are occurring and that the capacity to present self-peptides is not altered with time in healthy individuals. We call $\{\tau_n\}_{n\in\mathbb{N}}$ the stochastic process of infection times. The dynamics of cells between the events of the stochastic process are described in what follows.

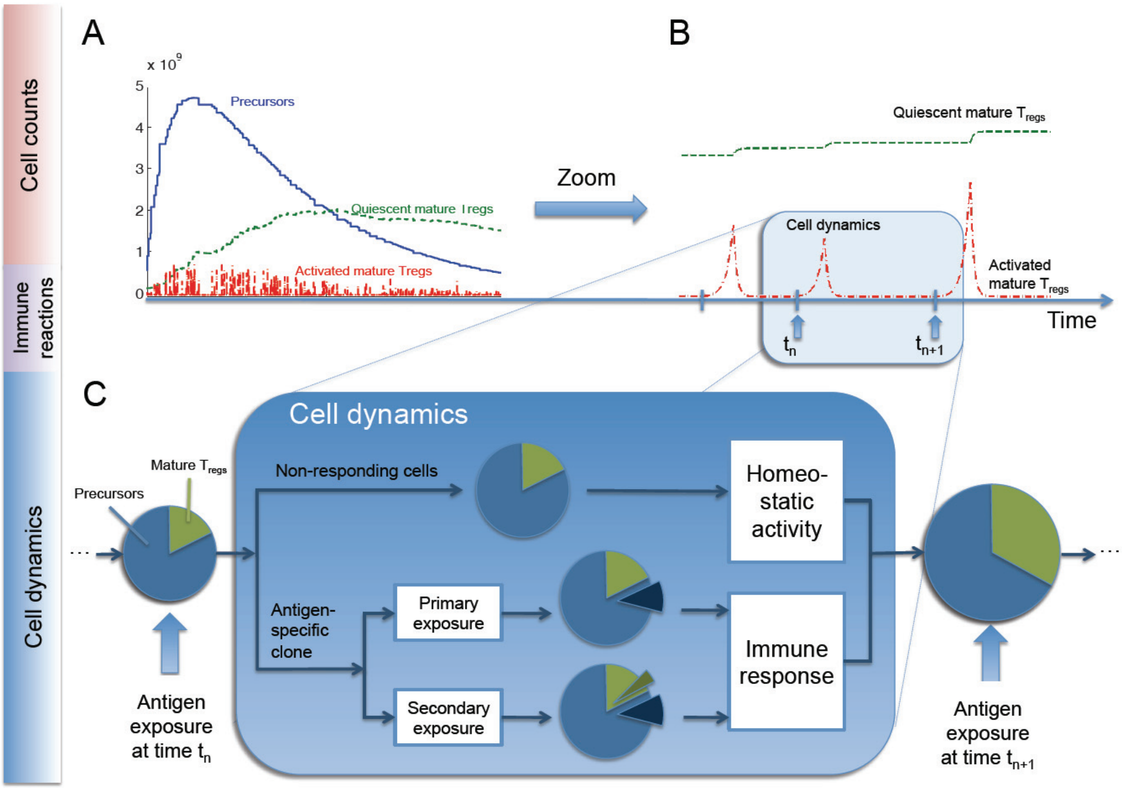


Figure 2.1: A sketch of the different components of the mathematical model. A: The output of our mathematical model are cell counts as function of time. We are considering three cell types, namely CD4$^+$CD25$^+$CD45RO$^-$ precursor T$_{\text{regs}}$ and CD4$^+$CD25$^+$CD45RO$^+$ quiescent or activated mature T$_{\text{regs}}$. B: The model is composed of a stochastic process generating the events corresponding to successful encounters with antigen-presenting cells leading to (auto-)immune responses. These responses are characterized by a transient increase in the activated mature T$_{\text{regs}}$ population. C: The dynamics of cells between two events of the stochastic process are described using ordinary differential equations. These equations take into account all events that affect the cell population size. A detailed description of each component of the cell dynamics can be found in the Generic Model section.

Cell dynamics Based on the expression of the surface protein CD45RO, one can sort two subpopulations of T$_{\text{regs}}$: CD4$^+$CD25$^+$FoxP3$^+$CD62L$^+$CD45RO$^-$ precursor and CD4$^+$CD25$^+$-FoxP3$^+$CD62L$^+$CD45RO$^+$ mature T$_{\text{regs}}$. For the sake of the mathematical model, we consider a third population that we call *activated mature* T$_{\text{regs}}$ and that has the same surface receptors as the *quiescent mature* T$_{\text{regs}}$ [2]. Activated mature T$_{\text{regs}}$ are composed of both precursors that experience peripheral antigens for the first time (they can be also activated precursors that have just acquired the mature phenotype), and mature T$_{\text{regs}}$ that are recruited into a secondary immune response. Thus, the mathematical model has three compartments: P, the precursor T$_{\text{regs}}$, Q, the quiescent mature T$_{\text{regs}}$, and R, the activated mature T$_{\text{regs}}$.

[2] Note that the population of activated T$_{\text{regs}}$ has been recently discovered as part of the peripheral in vivo pool of human T$_{\text{regs}}$ (Miyara et al., 2009b). We position our model with respect to this new perspective in Chapter 6.

a) Immune responses. *In vitro* experiments show that following acute antigenic stimulation, precursor $\mathrm{T_{regs}}$ activate and up-regulate the memory-type CD45RO receptor, as they loose their naive phenotype CD45RA receptor (Fritzsching et al., 2006; Valmori et al., 2005). In the meanwhile, they differentiate into mature $\mathrm{T_{regs}}$ able to exert their suppressive function. Once the antigenic stimulation is lost, a small proportion of all activated $\mathrm{T_{regs}}$ becomes long-lived mature cells, and the others die by apoptosis, similarly to other lymphocytes. As we are not modeling the dynamics of other cell types and of pathogens, we apply the two-phase immune response model used to describe $\mathrm{CD4^+CD25^-}$ and $\mathrm{CD8^+}$ cell dynamics (Althaus et al., 2007; De Boer et al., 2003, 2001; Fouchet and Regoes, 2008). We call τ_n the beginning of the n^{th} immune response to a foreign antigen ($n \in \mathbb{N}$) and $\tau_n + \delta$ the time at which the expansion phase ends. We assume that the effect of antigen on cells has a fixed duration δ, after which cells

Biological process	Assumptions	
(Auto-) Immune reactions process	1.	Major immune reactions modeled as a stochastic process with one (constant) parameter.
	2.	Minor immune reactions triggered by antigens sampled in the mucosa-associated lymphoid tissues included in the homeostasis activity.
	3.	Capacity to present self-peptides is not altered with age in healthy individuals.
	4.	During the fixed phase of an immune reaction, no other reactions are allowed.
Immune responses	5.	Two-phase immune response: a first phase of fixed length and a second phase of random length.
	6.	Exponential expansion/contraction of cells following immunogen stimulation.
	7.	10% of the expanded population of activated $\mathrm{T_{regs}}$ becomes long-lived mature $\mathrm{T_{regs}}$.
Homeostatic activity	8.	Unique source of precursors: thymus.
	9.	Four possible sources of mature $\mathrm{T_{regs}}$: thymus, differentiation of precursors, phenotype switching of effector $\mathrm{CD4^+CD25^-}$ T cells, and homeostatic proliferation.
	10.	Time-dependent thymic involution (exponential decline).
	11.	Constant rate of generation of $\mathrm{T_{regs}}$ from rapidly proliferating $\mathrm{CD4^+CD25^-}$ effector T cells under certain conditions.
Antigen specificity	12.	The size of the responding clone is chosen randomly (uniform distribution).
Primary / secondary infections	13.	The probability of primary infections is declining with age and is such that children experience a majority of primary infections and adults, a majority of secondary infections.

Table 2.1: Summary of all model assumptions.

stop their intense proliferation phase (Phase 1) and start dying by activation-induced cell death (Phase 2).

During Phase 1, precursor T$_{\text{regs}}$ may divide, die, or convert into mature effector cells at rate b. Quiescent mature T$_{\text{regs}}$ activate at rate f cells per day. Effector mature T$_{\text{regs}}$ may also proliferate and die. As the first phase is an expansion phase, we only consider a net population expansion rate, which should be interpreted as the cumulative effect of proliferation and death in the population of precursors (resp. activated mature T$_{\text{regs}}$). Thus, for parameter identification issues, we consider only one rate, called $a_P > 0$ (resp. $a_R > 0$), which is the net population expansion rate. The following differential equations express the dynamics of the expansion phase:

$$\begin{array}{rcl} \dot{P} & = & (a_P - b)P \\ \dot{Q} & = & -fQ \\ \dot{R} & = & bP + fQ + a_R R \end{array} \tag{2.1}$$

During Phase 2, precursor T$_{\text{regs}}$ die at rate d_P per day. Activated mature T$_{\text{regs}}$ die at rate d_R and convert to long-lived mature cells at rate c per day. The long-lived mature quiescent T$_{\text{regs}}$ have a slight decrease in their population size expressed by death rate d_Q. The differential equations corresponding to the contraction phase are the following:

$$\begin{array}{rcl} \dot{P} & = & -d_P P \\ \dot{Q} & = & -d_Q Q + cR \\ \dot{R} & = & -(c + d_R)R \end{array} \tag{2.2}$$

b) Homeostatic activity. The biological processes included in the homeostatic activity are the proliferation and death of cells for regulation of the population size, the death of cells because of their limited lifespan and the input of newly produced cells from the thymus of from another external source. We hypothesize two types of basal proliferation and death: a constant and density-dependent one. We call d' the constant death rate accounting for the limited lifespan of cells. We assume that a slow and steady cell division occurs at rate a'. Nevertheless, these constant renewal and death rates can not account for self-regulation of the cell population size. In a homeostatic situation, cell numbers are regulated by competition for limited resources, such as cytokines. This regulation can be achieved in three ways: density-dependent proliferation, density-dependent death or both. The exact way in which regulatory T cells perform their homeostatic regulation is currently unknown. It is however known that the *in vivo* homeostatic proliferation of murine natural T$_{\text{regs}}$ is not impaired by their anergic state (Gavin et al., 2002) and that this proliferation is involved in a feedback regulatory loop between dendritic cells and T$_{\text{regs}}$ (Darrasse-Jeze et al., 2009).

Because the homeostatic activity is an important issue to studying the lifelong dynamics of cells *in vivo*, we are considering all possible mechanisms that may have an effect on the model's outcome. Thus, we assume that homeostasis of precursors is achieved through density-dependent death at rate φ_P. For mature T$_{\text{regs}}$, we consider both a density-dependent death at rate φ_Q and a density-dependent Michaelis-Menten type proliferation at rate α, appropriate to make up for lymphopenic situations. Recent thymic emigrants enter both precursor and mature T$_{\text{regs}}$ populations at a time-dependent rate $g(t) = g_P(t) + g_Q(t)$. The constant term s_Q, added to the quiescent mature T$_{\text{regs}}$ population, represents the constant generation of regulatory T cells from rapidly proliferating CD4$^+$CD25$^-$ effector T cells under certain conditions (Akbar et al., 2007; Vukmanovic-Stejic et al., 2006). Remark that we only add this term to the quiescent mature T$_{\text{regs}}$ population, because we assume that the unique source of precursor CD4$^+$CD25$^+$CD45RO$^-$ T cells is the thymus. In addition, cells that are derived from rapidly proliferating CD4$^+$CD25$^-$ cells are antigen-experienced and thus have probably acquired the memory phenotype CD45RO before converting to FoxP3$^+$ regulatory cells. The differential equations describing the homeostatic activity are the following:

$$\begin{aligned}\dot{P} &= g_P(t) + (a'_P - d'_P)P - \varphi_P P^2 \\ \dot{Q} &= \underbrace{g_Q(t) + s_Q}_{\text{External contribution}} + \underbrace{(a'_Q - d'_Q)Q}_{\text{Density-independent regulation}} + \underbrace{\left(\frac{\alpha}{1 + Q/Q_M} - \varphi_Q Q\right)Q}_{\text{Density-dependent regulation}} \qquad (2.3) \\ \dot{R} &= 0,\end{aligned}$$

where Q_M is the size of the mature T$_{\text{regs}}$ population for which the homeostatic renewal of cells is half of the maximal rate α. Let $h_P = a'_P - d'_P$ and $h_Q = a'_Q - d'_Q$ be the cumulative effects of the constant renewal/death rates, $h_P \in \mathbb{R}$ and $h_Q \in \mathbb{R}$. Whenever negative, we will refer to these parameters as lifespan of precursors and mature T$_{\text{regs}}$. The thymic involution is represented as a decreasing exponential function of rate ν (Dutilh and De Boer, 2003; Marušić et al., 1998; Steinmann et al., 1985):

$$\begin{aligned}g_P(t) &= \sigma_P \exp(-\nu t) \\ g_Q(t) &= \sigma_Q \exp(-\nu t),\end{aligned}$$

where $\sigma_P = \sigma_0 * \%\text{CD25}_{\text{thymus}} * p_{\text{thymus}}$, $\sigma_Q = \sigma_0 * \%\text{CD25}_{\text{thymus}} * (1 - p_{\text{thymus}})$, $\%\text{CD25}_{\text{thymus}}$ is the percentage of CD25$^+$ cells inside thymic CD4$^+$ T cells, p_{thymus} is the percentage of precursors inside CD25$^+$cells that are output from the thymus, and σ_0 is the estimated thymic output of CD4$^+$ cells in a newborn.

c) Antigen specificity. To ease the notation in what follows, let $Y = (P, Q, R)'$. Antigen specificity is implemented in the following way. Eq. (2.1) and Eq. (2.3) are applied to a proportion π_n of the total number of T$_{\text{regs}}$, those representing the antigen-specific clone responding to the antigen that caused the immune reaction at time τ_n. We call this population $Y_{\text{clone}}(\tau_n)$. The other $1 - \pi_n$ proportion of cells do not participate in the n^{th} immune reaction and execute their homeostatic activity (Eq. (2.3)).

d) Primary/secondary infections. We take into account the difference between primary and secondary infections: when some antigen is encountered for the first time, no memory cells exist, but if the exposure is secondary, the organism already has memory cells associated to it at the time of exposure τ_n. We call $q(t)$ the probability that an infection at time t is primary and

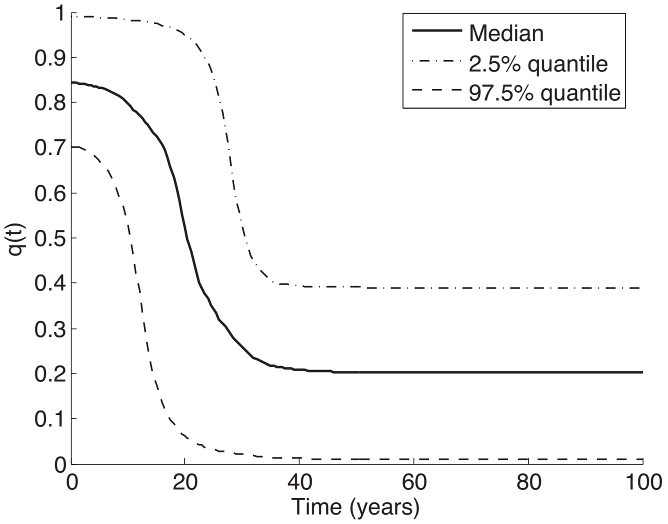


Figure 2.2: The probability $q(t)$ of a primary infection is decreasing with age. An estimation of the distribution for $q(t)$ defined in Eq.(4) with parameters sampled from the *a priori* distributions of Table 2.4.

does not involve mature T$_{\text{regs}}$. Thus, with probability $q(\tau_n)$, the responding clone at time τ_n is set to $Y_{\text{clone}}(\tau_n) = \pi_n(P(\tau_n), 0, 0)'$. Otherwise, with probability $1 - q(\tau_n)$, the responding clone is set to $Y_{\text{clone}}(\tau_n) = \pi_n(P(\tau_n), Q(\tau_n), 0)'$. Intuitively, the unexperienced immune system of young individuals is confronting more primary infections than adults. We therefore define the following sigmoid function describing phenomenologically the time-dependence of parameter q:

$$q(t) = \frac{K_1 - K_2}{1 + \exp(\omega(t - t_h))} + K_2 \tag{2.4}$$

where K_1 is the (approximate) proportion of primary infections at birth, K_2 is the limit proportion of primary infections at adulthood, ω is the maximum decline rate and t_h is the age at which the proportion of primary infections is $q(t_h) = (K_1 + K_2)/2$. Note that contrary to CD4⁺CD25⁻ memory T cells, mature T$_{\text{regs}}$ require additional conditions for their activation at the time of a secondary antigen exposure. Because these cells are non-proliferating and do not produce the growth factor IL-2 themselves, they need optimal stimulation conditions and a high concentration of IL-2 in order to initiate a response (Hombach et al., 2007). All this is taken into account in the above definition of $q(t)$.

In order to eliminate as much as possible any dependence of the final results on the exact shape of $q(t)$, we have defined random distributions on the parameters K_1, K_2, ω and t_h (see Table 2.4).

Model scenarios

The model scenarios are obtained from the generic model by applying a set of constraints to the following parameters: $g_Q(t), s_Q, \varphi_Q, \alpha, a_P, a_R$ and d_P. From the homeostasis dynamics of the generic model, we define three *homeostasis scenarios*:

(i) No homeostatic regulation of mature T$_{\text{regs}}$: due to their anergic state, the only observed phenomena is the slow and steady density-independent proliferation and death ($g_Q(t) = s_Q = \varphi_Q = \alpha = 0$);

(ii) Homeostatic regulation of mature $\mathrm{T_{regs}}$: in addition to the constant proliferation and death rates, we allow for density-dependent regulation mechanisms of mature $\mathrm{T_{regs}}$. As we search for the minimal model able to explain the data, we consider two mutually exclusive sub-settings:
  (a) density-dependent death of mature $\mathrm{T_{regs}}$ ($\varphi_Q > 0$, $\alpha = 0$);
  (b) density-dependent proliferation of mature $\mathrm{T_{regs}}$ ($\varphi_Q = 0$, $\alpha > 0$).

(iii) External input: we consider two external contributions to the mature $\mathrm{T_{regs}}$ population ($\alpha = 0$):
  (a) thymic output of mature $\mathrm{T_{regs}}$ Vanhecke et al. (1995) ($g_Q(t) > 0$);
  (b) peripheral differentiation of CD4$^+$CD25$^-$ non-regulatory cells into their regulatory counterpart Vukmanovic-Stejic et al. (2006) ($s_Q > 0$).

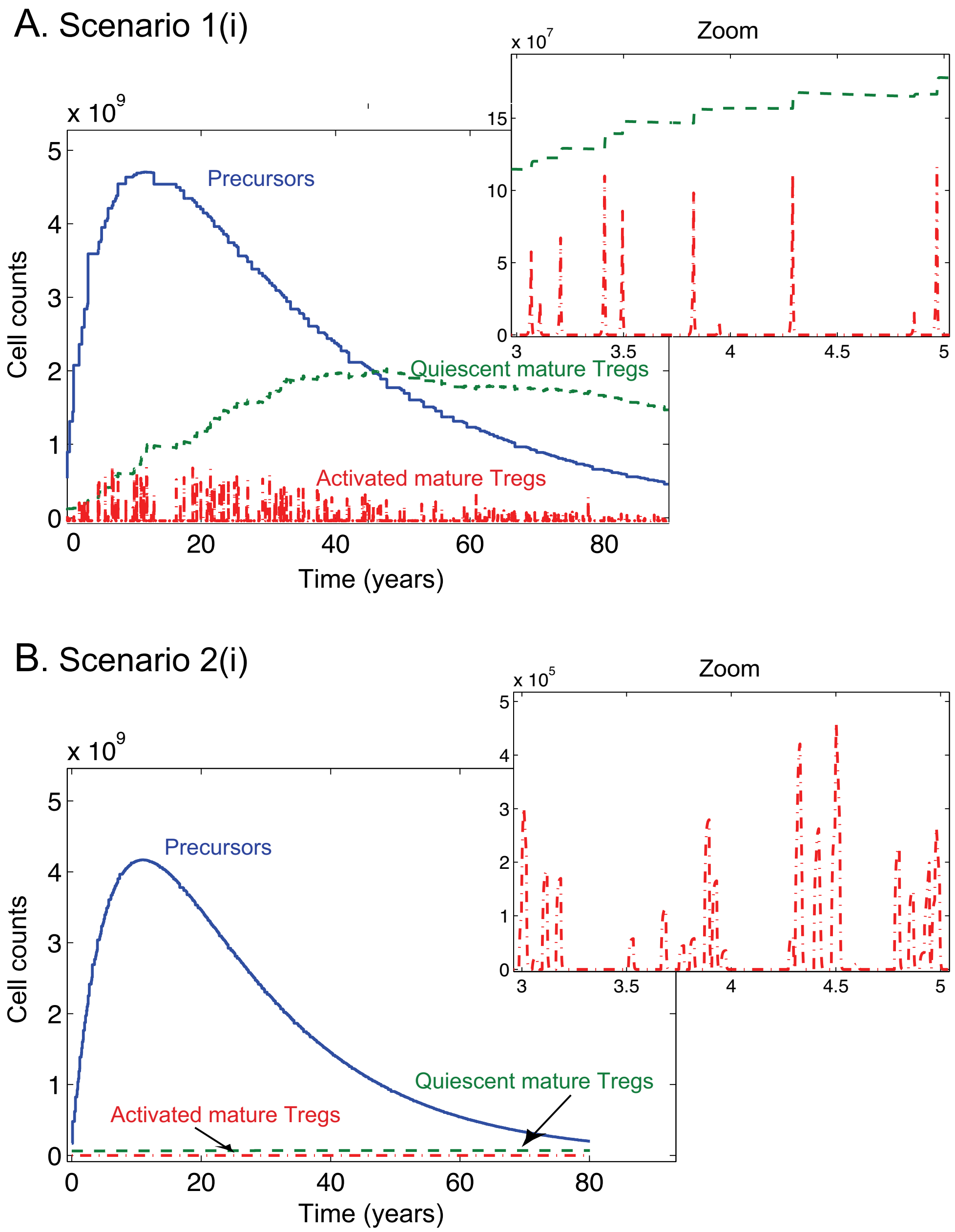
Homeostasis scenarios		Parameter setting
(i)	No homeostasis	$g_Q(t) = 0, s_Q = 0, \varphi_Q = 0, \alpha = 0$
(ii)	Homeostatic regulation	
	(a) Density-dependent death	$g_Q(t) = 0, s_Q = 0, \varphi_Q > 0, \alpha = 0$
	(b) Density-dependent proliferation	$g_Q(t) = 0, s_Q = 0, \varphi_Q = 0, \alpha > 0$
(iii)	External input	
	(a) From thymus	$g_Q(t) > 0, s_Q = 0, \varphi_Q = 0, \alpha = 0$
	(b) Peripheral differentiation of CD4+CD25- T cells	$g_Q(t) = 0, s_Q > 0, \varphi_Q = 0, \alpha = 0$

Table 2.2: Definition of the homeostasis scenarios. Parameter meanings: $g_Q(t)$, time-dependent input of recent thymic emigrants into the peripheral mature $\mathrm{T_{regs}}$ population; s_Q, constant input of regulatory T cells from rapidly proliferating CD4$^+$CD25$^+$ effector T cells, φ_Q and α are the density-dependent death and proliferation rates.

From the cell dynamics in response to self and foreign antigens, we define four *proliferation scenarios* accounting for different proliferation capacities of $\mathrm{T_{regs}}$:

1) Both precursor and activated-mature $\mathrm{T_{regs}}$ proliferate and die in response to an antigen stimulus;
2) Neither precursor nor mature $\mathrm{T_{regs}}$ proliferate, but precursors are very resistant to apoptosis, so they do not die during an immune response;
3) Precursors proliferate in response to antigen, but as soon as they differentiate into CD45RO$^+$ mature $\mathrm{T_{regs}}$, they stop proliferating and die extensively because of the high levels of expression of the CD95 receptor Fritzsching et al. (2006) and because of the downregulation of the anti-apoptotic protein Bcl-2 Yamaguchi et al. (2007);
4) Precursors do not proliferate when they are CD45RO$^-$, but the proliferation starts while they acquire the mature profile.

Table 2.2 and Table 2.3 summarize the above settings. We construct a *model scenario* by choosing one homeostasis and one proliferation scenario, thus obtaining 20 model scenarios, referred from now on as scenario 1(i), 2(i), ..., 4(i), 1(ii)a, ..., 4(ii)b, 1(iii)a, ..., 4(iii)b.



A. Scenario 1(i)

B. Scenario 2(i)

Figure 2.3: Typical model trajectories for scenarios 1(i) and 2(i). The solid line represents the evolution of precursors, the dashed one the dynamics of quiescent mature $\mathrm{T_{regs}}$, and the dash-dotted line represents the activated mature $\mathrm{T_{regs}}$. Each spike of the activated mature $\mathrm{T_{regs}}$ corresponds to a major immune reaction (an event of the stochastic process). The ratio inversion observed in A is an example of slow accumulation of mature $\mathrm{T_{regs}}$ following an antigen stimulation. Parameter values: average values of the distributions of Table 2.4 with the following values for the fitted parameters: $h_P = -5 \times 10^{-4}$, $h_Q = -5 \times 10^{-5}$, $\lambda = 130$ for scenario 1(i), and $h_P = -5 \times 10^{-4}$, $h_Q = 0$, $\lambda = 30$ for scenario 2(i).

Proliferation scenarios	Proliferation capacity of...		Parameter setting
	Precursors	Mature T$_{\text{regs}}$	
1	Proliferate	Proliferate	$a_P > 0, a_R > 0, d_P > 0$
2	Do not proliferate	Do not proliferate	$a_P = 0, a_R = 0, d_P = 0$
3	Proliferate	Do not proliferate	$a_P > 0, a_R = 0, d_P > 0$
4	Do not proliferate	Proliferate	$a_P = 0, a_R > 0, d_P = 0$

Table 2.3: Definition of the four proliferation scenarios. Parameter meanings: a_P, antigen-induced proliferation rate of precursors; a_R, antigen-induced proliferation rate of mature T$_{\text{regs}}$; d_P, antigen-induced death rate of precursors.

The model at a glance: typical model trajectories

In order to gain some insight into the model dynamics, we present in this section an example of execution of our model. Figure 2.3 shows the number of precursors (solid line) and mature T$_{\text{regs}}$ (resting: dashed line and activated: dash-dotted line) as function of time in human peripheral's blood. The zoomed sections allow a closer look at the activated T$_{\text{regs}}$ population which is mainly present during major immune reactions, events of the stochastic process.

In Figure 2.3A is depicted a typical trajectory for scenario 1(i) where the ratio between precursors and mature T$_{\text{regs}}$ is inverted in the early years of adulthood, whereas in Figure 2.3B is shown a typical trajectory for scenario 2(i) where the ratio inversion is not observed. All model scenarios will be evaluated based on their capacity to reproduce this ratio inversion and on the capacity to reproduce the total amount of T$_{\text{regs}}$ suggested by the data. Note that because the thymus produces thymocytes with a constant proportion of each subtype of T$_{\text{regs}}$, there are four means of achieving a precursor/mature T$_{\text{regs}}$ ratio inversion: 1) a difference in lifespans of precursors and mature T$_{\text{regs}}$; 2) a difference in the homeostatic proliferation rates of precursors and mature T$_{\text{regs}}$; 3) the slow accumulation of mature T$_{\text{regs}}$ following an antigen stimulation; 4) an external source of mature T$_{\text{regs}}$. The first mean is present in all model scenarios and it acts in combination with the other three means. Thus in each model scenario, we observe a combination of the above means and this combination can be sufficient or not to achieve the ratio inversion. For example, in Figure 2.3 where there is no homeostatic regulation, the lifespan of precursors is larger than the lifespan of mature T$_{\text{regs}}$ ($|h_P| > |h_Q|$), but this lifespan difference is not sufficient to achieve the ratio inversion. Here, the slow accumulation of mature T$_{\text{regs}}$, present in scenario 1(i) and not in scenario 2(i), is needed.

Parameters

In order to cope with the large number of parameters and to decrease the effect of single values on the dynamics of the model, we split the parameters in two groups: G_1, parameters for which some *a priori* information is found, and G_2, unknown parameters that must be fitted to the data (see Section 2.2.2 for a detailed description of the data). The probability distribution of parameters of group G_1 is defined in the following way: we find in the literature either a parameter together with a confidence interval accounting for the uncertainty of the estimation, or a finite range without any preferred value specified. In the former case, we use a Gaussian distribution with mean and variance given by the confidence interval. In the latter case, we use a uniform distribution over the range.

Our *in vitro* assays have shown that the expansion rate a_P of T$_{\text{regs}}$ precursors is similar to the one of CD4$^+$CD25$^-$ naive T cells and that the expansion rate a_Q of activated mature T$_{\text{regs}}$ is 1.4 to 2 times lower than the expansion rate of precursors. To our knowledge, the *in vivo* expansion rate of precursors (or even of CD4$^+$CD25$^-$ naive T cells) in humans has not yet been measured. The *in vivo* T cell responses in mice have been quantified in (De Boer et al., 2003) which gives a first idea about the order of magnitude of parameters such as the proliferation and death rates of CD4$^+$ T cells, as well as the duration of a T cell response. For a_P, we find in (De Boer et al., 2003) values in the range 1 - 1.7 per day, so we are sampling from an uniform distribution between 1 and 1.7 (day^{-1}) ($a_P \sim \mathcal{U}(1, 1.7)$). We set $a_R = a_P/K$, where K takes two different values, 1.4 and 5, the latter being an extreme value representing the case where mature T$_{\text{regs}}$ have a very limited proliferating capacity compared to precursors. The value of the differentiation rate b of precursors into mature T$_{\text{regs}}$ defines two settings: in the first, $b \sim \mathcal{U}(1, 10)$ (Burroughs et al., 2006) and in the second, we apply the constraint $b \leq a_P$, thus $b \sim \mathcal{U}(0.5, 1)$. We will see that these settings give slightly different results. The activation rate of quiescent mature T$_{\text{regs}}$ is set to half of the differentiation rate of precursors, i.e., $f = b/2$, because mature T$_{\text{regs}}$ are anergic and thus slow to activate. The duration δ_n of the expansion phase is drawn from a Gaussian distribution $\mathcal{N}(7.65, 0.72)$ (*in vitro* assay and (De Boer et al., 2003)). Both death rates d_P and d_R are considered as uniform $\mathcal{U}(1, 2)$ because they should have similar or higher values to the proliferation rates for system stability issues. It is known that $\sim 5 - 10\%$ of all effector cells become long-lived memory cells. Thus, we set $c = d_R/9$, which corresponds to 10% of the maximum value of the effector response. The death rate d_Q is distributed according to $\mathcal{N}(0.0013, 0.0002)$ (De Boer et al., 2003).

The thymic input is calculated as follows. For the value of the total number of CD4$^+$ T cells output from the thymus of a newborn (σ_0) and for the thymic involution rate (ν), we found the following parameter settings in literature: $\sigma_0 = 1.98 \times 10^8$ cells per day and $\nu = 0.024$ per year (Marušić et al., 1998; Steinmann et al., 1985); $\sigma_0 = 4.48 \times 10^8$ cells per day and $\nu = 0.05$ per year (Clark et al., 1999; Dutilh and De Boer, 2003). We therefore use an uniform distribution $\mathcal{U}(1{\times}10^8, 5{\times}10^8)$ for σ_0 and $\mathcal{U}(0.01, 0.06)$ for ν to include the above settings in our simulations. The percentage of CD25$^+$ cells among all CD4$^+$ thymocytes ($\%\text{CD25}_{\text{thymus}}$) was measured to be 1.3% in young children (unpublished data).

The proportion π_n, $n \in \mathbb{N}$, of cells that respond to a particular antigen is drawn uniformly between 10^{-7} and 10^{-4} cells (De Boer and Perelson, 1993). To estimate the total number of T cells in human's peripheral blood as a function of age, $T(t)$, we use the fact that the mean blood volume of a healthy 70 kg individual is 5 liters (Feldschuh and Enson, 1977) and it contains about 10^{11} T cells (Clark et al., 1999). We then use a function that maps age to the mean weight of individuals (CDC, 1988–1994; WHO, 2009).

The constant and density-dependent homeostatic proliferation and death rates (h_P, h_Q, φ_P, φ_Q, α, Q_M), as well as the infections occurrence rate (λ) and the external input from the thymus (p_{thymus}) or from CD25$^-$ conversion (s), are fitted to the data. The time-unit of the model is 1 day. Table 2.4 summarizes all parameter values and distributions.

	Meaning (units)	Value or Distribution	Source
1.	Expansion of precursors (1/day)	$a_P \sim \mathcal{U}(1, 1.7)$	Biological assumption, De Boer et al. (2003)
2.	Expansion of activated mature T_regs (1/day)	$a_R = a_P/1.4$ or $a_R = a_P/5$	*In vitro* assay, Burroughs et al. (2006)
3.	Differentiation of precursors (1/day)	$b \sim \mathcal{U}(1, 10)$ or $\mathcal{U}(0.5, 1)$	Burroughs et al. (2006)
4.	Resting of mature T_regs (1/day)	$c = d_R/9$	10% of response peak
5.	Death of precursors (1/day)	$d_P \sim \mathcal{U}(1, 2)$	Similar to a_P
6.	Death of activated mature T_regs (1/day)	$d_R \sim \mathcal{U}(1, 2)$	Similar to a_P
7.	Death of quiescent mature T_regs (1/day)	$d_Q \sim \mathcal{N}(0.0013, 0.0002)$	De Boer et al. (2003)
8.	Activation of mature T_regs (1/day)	$f \sim \mathcal{U}(0.5, 5)$ or $f \sim \mathcal{U}(0.25, 0.5)$	Biological assumption
9.	Duration of expansion phase (days)	$\delta \sim \mathcal{N}(8, 1.33)$	*In vitro* assay
10.	Antigen specificity	$\pi_n \sim \mathcal{U}(10^{-7}, 10^{-4})$	De Boer and Perelson (1993)
11.	Probability of primary infections at birth	$K_1 \sim \mathcal{U}(0.7, 1)$	Biological assumption
12.	Limit proba. of primary infections at adulthood	$K_2 \sim \mathcal{U}(0, 0.4)$	Biological assumption
13.	Age at which the proba. of primary infections is half (days)	$t_h \sim \mathcal{U}(10 * 365, 30 * 365)$	Biological assumption
14.	Maximum decline rate of $q(t)$	$\omega \sim \mathcal{U}(0.01, 0.5)$	Biological assumption
15.	Mean blood volume in a healthy 70kg individual (liters)	5	Feldschuh and Enson (1977)
16.	Number of T cells in 5l of blood (cells)	$T_{5l} \sim \mathcal{U}(1 \times 10^{11}, 5 \times 10^{11})$	Clark et al. (1999)
17.	Number of CD4^+ cells output from the thymus of a newborn (cells/day)	$\sigma_0 \sim \mathcal{U}(1 \times 10^8, 5 \times 10^8)$	Clark et al. (1999); Dutilh and De Boer (2003); Marušić et al. (1998); Steinmann et al. (1985)
18.	Thymic involution rate (1/day)	$\nu \sim \mathcal{U}(0.01/365, 0.06/365)$	Clark et al. (1999); Dutilh and De Boer (2003); Marušić et al. (1998); Steinmann et al. (1985)
19.	% CD25^+ inside CD4^+ cells output from the thymus	$\%\text{CD25}_\text{thymus} = 0.0065$	Data
20.	% CD25^+ among all T cells in cord blood	$\%\text{CD25}_\text{cb} \sim \mathcal{N}(0.03, 0.01)$	Data
21.	Proportion of precursors inside $\text{CD4}^+\text{CD25}^+$ in cord blood	$p_\text{cb} \sim \mathcal{N}(0.78, 0.064)$	Data
22.	Proportion of mature T_regs inside $\text{CD4}^+\text{CD25}^+$ in cord blood	$q_\text{cb} \sim \mathcal{N}(0.21, 0.064)$	Data
23.	Number of precursors at birth (cells)	$P_0 = T(0) * \%\text{CD25}_\text{cb} * p_\text{cb}$	Data
24.	Number of mature T_regs at birth (cells)	$Q_0 = T(0) * \%\text{CD25}_\text{cb} * q_\text{cb}$	Data

Table 2.4: *A priori* parameter distributions (parameters of group G_1). $\mathcal{N}(\mu, \sigma)$ stands for a Gaussian distribution with mean μ and standard deviation σ and $\mathcal{U}(a, b)$ stands for a uniform distribution on [a, b]. Parameters h_P, h_Q, λ, φ_P, φ_Q, Q_M, α, p_thymus and s are fitted to the data.

2.2.2 Immunological data

Biological specimens

A transversal study was performed on the peripheral blood of 120 healthy subjects constituted by 60 males with a median age of 52.6 (range 20-81) and 60 females with a median age of 49.32 (range 19-78). Samples of peripheral blood were obtained from laboratory coworkers or from the Blood Bank of the Centre Hospitalier Universitaire Vaudois, University of Lausanne, while the 7 samples of cord blood and the 2 thymuses were obtained from the Department of Clinical Chemistry, Microbiology and Immunology, University Hospital, University of Ghent, Belgium. Cord and peripheral blood mononuclear cells were isolated using standard Ficoll-Hypaque (Amersham Pharmacia Biotech, Piscataway, NJ) gradients centrifugation. Blood specimens were collected under protocols approved by the Institutional Review Boards of the above mentioned Institutions.

FACS analysis and sorting

Cell surface analysis and sorting were performed using a combination of a panel of surface markers: PercP or PercP Cy5.5 conjugated mouse anti-human CD4 (Becton Dickinson, Franklin, NJ), APC-conjugated mouse anti-human CD25 (BD PharMingen, San Diego, CA), PE-conjugated mouse anti-human CD62L or PE-conjugated mouse anti-human CD71 (Becton Dickinson, Franklin, NJ), FITC or PE-conjugated mouse anti-human CD45RO (BD PharMingen, San Diego, CA) and FITC Annexin-V (Becton Dickinson, Franklin, NJ). For cell-sorting experiments CD4$^+$CD25$^+$CD62L$^+$CD45RO$^-$, CD4$^+$CD25$^+$CD62L$^+$CD45RO$^+$, CD4$^+$CD25$^-$-CD62L$^+$CD45RO$^-$ and CD4$^+$CD25$^-$CD62L$^+$CD45RO$^+$ cell populations were isolated from the peripheral blood. The grade of cell purity in all the sorting experiments was more than 97%. All flow cytometric analyses were performed on a FACS Calibur and cell sorting on a FACS Aria (Becton Dickinson Systems, Franklin, NJ). For intracellular FOXP3 analysis, cell preparations were fixed and permeabilized with fixation/permeabilization buffers (eBioscience) after staining of cell surface markers and stained with FITC-conjugated rat antiñhuman FOXP3 (eBioscience).

Regulatory T cells have been sorted and purified from peripheral and cord blood based on a CD4$^+$CD25$^{\mathrm{high}}$CD62L$^+$ gate. In order to analyze the purity of naturally occurring T$_{\mathrm{regs}}$, the above population has been stained for Foxp3. It is clear from Figure 2.4 that the cells sorted with this method are in majority Foxp3 positive. CD4$^+$CD25$^-$CD62L$^+$ cells have been sorted and stained in the same way. The majority of all CD25$^-$ cells are Foxp3 negative. Naturally occurring T$_{\mathrm{regs}}$ express CD25 constitutively and contrary to recently activated CD25$^-$ T cells, they do not downregulate this receptor. It is not excluded that the above sorting from peripheral blood involves some adaptive T$_{\mathrm{regs}}$ which are also included in the mathematical model presented hereafter.

Suppression assay

Sorted purified CD4$^+$CD25$^-$ T cells have been stimulated *in vitro* with anti-CD3 and anti-CD28 antibodies for 3 days alone (positive control) or in presence of sorted purified CD4$^+$-

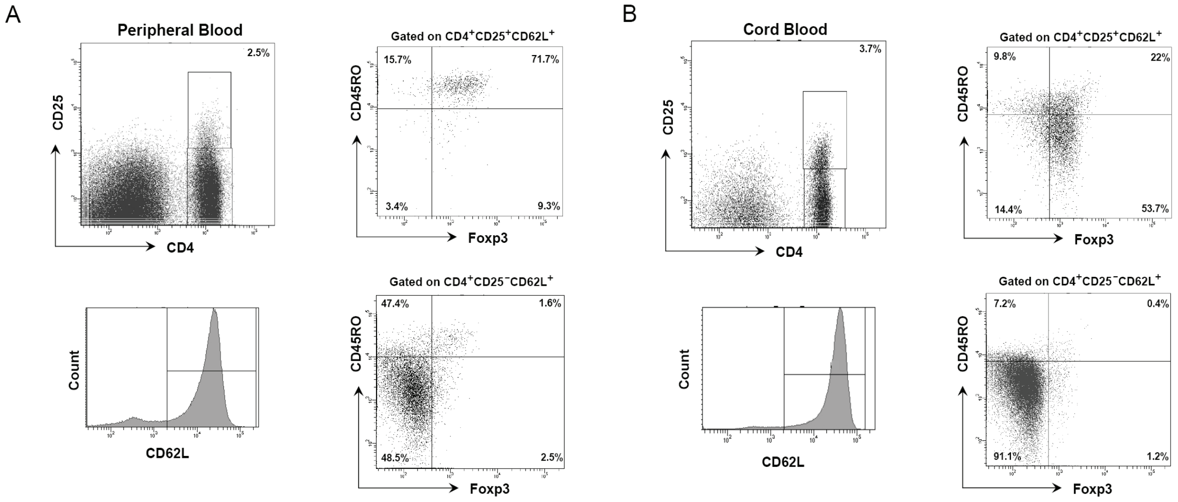


Figure 2.4: Isolation and characterization of human CD4$^+$CD25$^+$ T$_{\text{regs}}$ from peripheral and cord blood. A. Representative flow cytometry profiles of one peripheral blood sample (1 out of 120). B. Representative flow cytometry profiles of one cord blood sample (1 out of 7). Regulatory T cells have been sorted and purified from peripheral and cord blood based on the gate on CD4$^+$CD25$^{\text{high}}$CD62L$^+$ cells; the gate for non-regulatory T cells was on CD4$^+$CD25$^{\text{neg}}$CD62L$^+$ cells. In both peripheral and cord blood, the majority of CD4$^+$CD25$^+$ cells are Foxp3$^+$ and the majority of CD4$^+$CD25$^-$ cells are Foxp3$^-$.

CD25$^+$CD62L$^+$CD45RO$^+$ T cells to test the suppressive function of the latter on the proliferation capacity of the former population. CD4$^+$CD25$^+$CD62L$^+$CD45RO$^+$ T cells have been added directly in the same well or in a transwell to test if their suppressive function needs cell-to-cell contact or is mediated by cytokines. CD4$^+$CD25$^+$CD62L$^+$CD45RO$^+$ T cells were able to strongly suppress (96%) the proliferation of CD4$^+$CD25^{neg} T cells when co-cultured in the same well, but lost all suppressive activity once cultured in a transwell (Figure 2.5).

Proliferation, activation and differentiation experiments

In order to assess the proliferation capacity, peripheral blood sorted cell populations (CD4$^+$-CD25$^+$CD62L$^+$CD45RO$^-$, CD4$^+$CD25$^+$CD62L$^+$CD45RO$^+$, CD4$^+$CD25$^-$CD62L$^+$CD45RO$^-$ and CD4$^+$CD25$^-$CD62L$^+$CD45RO$^+$) were plated at 5×10^3 cells per well in 96 U-bottomed plates in RPMI 1640 plus 10% FBS in the presence of 5×10^4 irradiated (4000 rads) syngeneic peripheral blood mononuclear cells. Cells were cultured in the presence of soluble anti-CD3 (OKT3) plus soluble anti-CD28 (BD Pharmingen, San Diego, CA) with or without IL-2. The proliferation kinetic was followed by FACS over 15 days and by [3H]-Thymidine incorporation in correspondence of proliferative peaks. The activation and differentiation were monitored during 15 days by the expression of activation markers such as CD25 and CD71 (Transferrin Receptor) and the marker of mature cells CD45RO. Cell death by apoptosis was measured in parallel for each population by expression of Annexin-V.

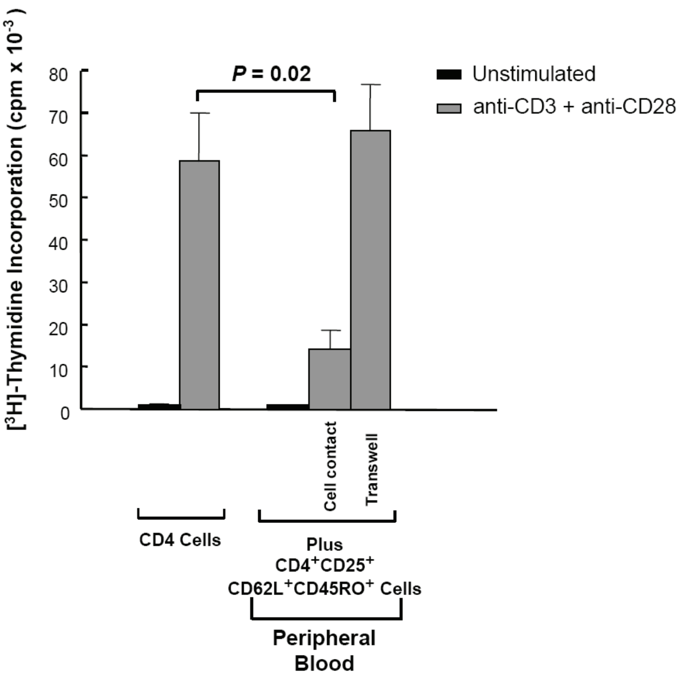


Figure 2.5: Cell-to-cell contact is needed by the CD4$^+$CD25$^+$CD62L$^+$CD45RO$^+$ T cells to exert their suppressive function, while cytokines do not seem to play any role.

Extrapolation of a mathematical model from empirical observations

Prior to formulating a mathematical model, we studied the proliferation kinetics of three populations (CD4$^+$CD25$^+$CD62L$^+$CD45RO$^-$, CD4$^+$CD25$^+$CD62L$^+$CD45RO$^+$, and CD4$^+$CD25$^-$-CD62L$^+$CD45RO$^-$ cells) that may be involved in the maintenance of a constant pool of regulatory T cells. The assay was performed several times using sorted cells from different subjects in order to extrapolate some parameter values for the model, free from the individual vices. Each of the sorted cell populations was cultivated in presence of a polyclonal stimulus in order to stimulate all the clones contained in a population. All the experiments were designed in such a way that IL-2 is added at the beginning of the culture and every two days, because of the limited proliferation capacity of mature T$_{\text{regs}}$. In fact, these cells did not show any appreciable proliferation in other parallel experiments driven in the absence of IL-2. Each experiment lasted 15 days to permit the estimation of parameter values such as the duration of the proliferation phase, the amount of cells produced, the number of cells that were activated (CD71$^+$), differentiated and acquired a mature phenotype (CD45RO$^+$), the cell death (Annexin$^+$ cells) and the number of surviving cells. We then studied retrospectively the distribution of the above populations in the peripheral blood of 120 healthy individuals, 7 cord blood samples and 2 thymuses, obtained from two children of 1 and 3 years old, who underwent cardiac surgery. In this way, we obtained the distribution of the above populations as a function of age.

2.2.3 Model evaluation procedure and parameter fitting

The model assessment procedure consists in the following steps: first, we consider the mean model dynamics and we fit the unknown parameters to the biological data using a least-squares procedure (implemented in Matlab). The mean model is derived from the stochastic one by fixing the parameters of group G_1 to the mean value of their *a priori* distribution and by averaging over the remaining random quantities. The latter are Y_{clone}, the size of the responding clone, and $\Delta_n = \tau_{n+1} - \tau_n$, the time-interval between two consecutive immune reactions. Based on this fit, we can reject some model scenarios. However, this result may depend on the mean values of the *a priori* distributions. In order to eliminate this possibility, we fit the stochastic model (in which parameters of group G_1 are random variables) using a Bayesian approach. This fit consists in maximizing the likelihood function, a measure of how good is the model in explaining the data. It is the probability that, given the best-fit parameters and the *a priori* distributions, the data is a realization of a particular model scenario. The higher this probability, the closer the model is to the data. The likelihood of a model scenario is computed as follows.

Consequently to the fact that the parameters of group G_1 are defined as random variables, $\{P(t)\}_{t\in\mathbb{R}}$, $\{Q(t)\}_{t\in\mathbb{R}}$ and $\{R(t)\}_{t\in\mathbb{R}}$ are stochastic processes. In order to fit each model scenario to the biological data, we transform the above processes in the format of the data, i.e., in terms of the ratio precursor/mature T_{regs} and of the percentage of $\text{CD4}^+\text{CD25}^+$ cells inside all T cells, in logarithmic scale. Thus, we consider two other stochastic processes, namely $\{U(t)\}_{t\in\mathbb{R}}$ and $\{V(t)\}_{t\in\mathbb{R}}$, given by

$$\begin{aligned} U(t) &:= \log\left(\frac{P(t)}{Q(t)+R(t)}\right) \\ V(t) &:= \log\left(\frac{P(t)+Q(t)+R(t)}{T(t)}\right), \end{aligned}$$

where $T(t)$ is the total number of T cells at time t.

Let $g_{UV}(u,v;t)$ be the joint density of U and V. Because it is difficult to compute $g_{UV}(u,v;t)$ analytically, we estimate it using Monte Carlo simulation, as follows. The model is executed M times for a time horizon of 85 years with parameters drawn according to the distributions of Table 2.4. At each time, the density of trajectories in the space of U and V is estimated using a bivariate histogram. Call $\hat{g}^{(s)}_{UV}(u,v;t)$ the estimated density of model scenario s at time t. Using the M model replicates, we construct a confidence interval for the bivariate histogram Davison (2003). We then consider the relative size of the confidence interval of each bin. If the latter is greater than 10% for some bin, the number of replicates M is increased. The density estimation is considered satisfactory if the relative size of the confidence interval of all bins is smaller than 10%. In the case of our model scenarios, M is taking values between 100 000 (in most cases) and 500 000.

Each model scenario s is fitted to the data by maximizing the log-likelihood function defined as

$$\ell_s(\theta) = \sum_{i=1}^{N} \log\left(\hat{g}^{(s)}_{UV}(u_i, v_i; \theta | t_i)\right),\ \theta \in \Theta_s,$$

where N is the number of data points ($N = 126$) and $\hat{g}^{(s)}_{UV}(u_i, v_i; \theta | t_i)$ is regarded as a function of θ for (u_i, v_i) and t_i fixed. The parameter space Θ_s is defined by the values that can take the free parameters $\theta = (h_P,\ h_Q,\ \lambda,\ \varphi_P,\ \varphi_Q,\ Q_M,\ \alpha,\ p_{\text{thymus}},\ s)$. See Figure 2.6 for a visual explanation of the likelihood computation.

The log-likelihood of is also used for objectively comparing the different model scenarios by performing a likelihood ratio test (F-test) for nested models that formally rejects the scenar-

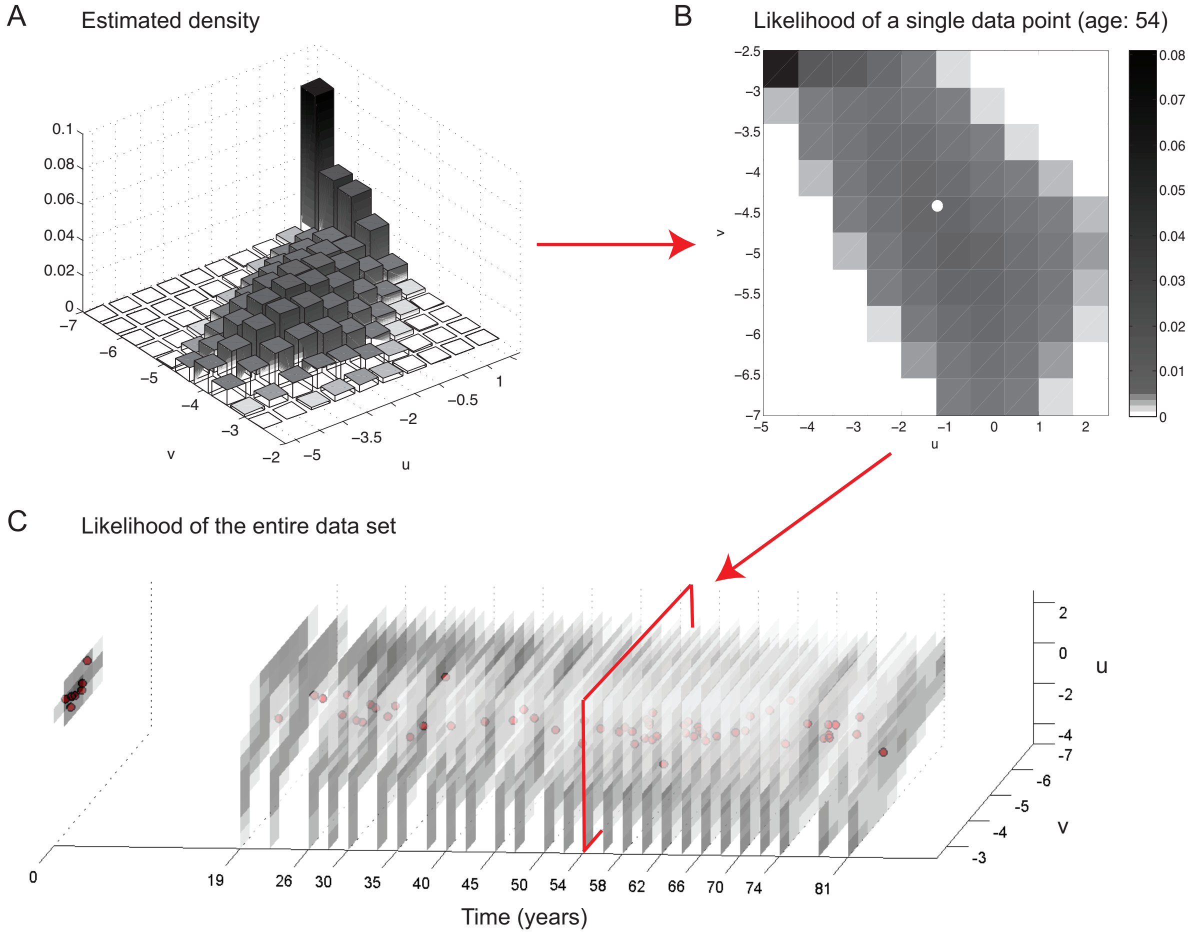


Figure 2.6: Likelihood computation under model scenario s. A: at each age t_i, the density $\hat{g}_{UV}^{(s)}(u, v; t_i)$ of the model is estimated. u stands for the ratio precursor/mature T$_{\text{regs}}$ and v, for the percentage of CD4$^+$CD25$^+$ cells inside all T cells. B: the estimated density $\hat{g}_{UV}(u, v; \theta|t_i)$ of the model is used to compute the likelihood of each data point. In this example, the likelihood of the point $(u, v) = (-1.21, -4.41)$ (age: 54 years) is equal to 0.028. C: the log-likelihood of the entire data set is equal to the sum of the log-likelihoods of all data points.

ios that are not compatible with the data. Two models are nested if the parameter space of one of them is a subset of the parameter space of the other. Proliferation scenarios 2, 3 and 4 (null hypothesis) are subsets of scenario 1 and homeostasis scenario (i) (null hypothesis) is a subset of all other homeostasis scenarios. The null hypothesis is rejected if the p-value of the test is < 0.01.

2.3 Results

2.3.1 Properties of T_{regs} dynamics

The immunological data are presented in Figure 2.7, in which are plotted A: the percentage of $CD4^{+}CD25^{+}$ cells inside all T cells, B: the ratio $CD4^{+}CD25^{+}CD45RO^{-}$ precursor / $CD4^{+}$-$CD25^{+}CD45RO^{+}$ mature T_{regs} as a function of age and C: the proportions of precursor and mature T_{regs} inside the $CD4^{+}CD25^{+}$ compartment. A direct analysis of the above data sets allows the identification of two properties.

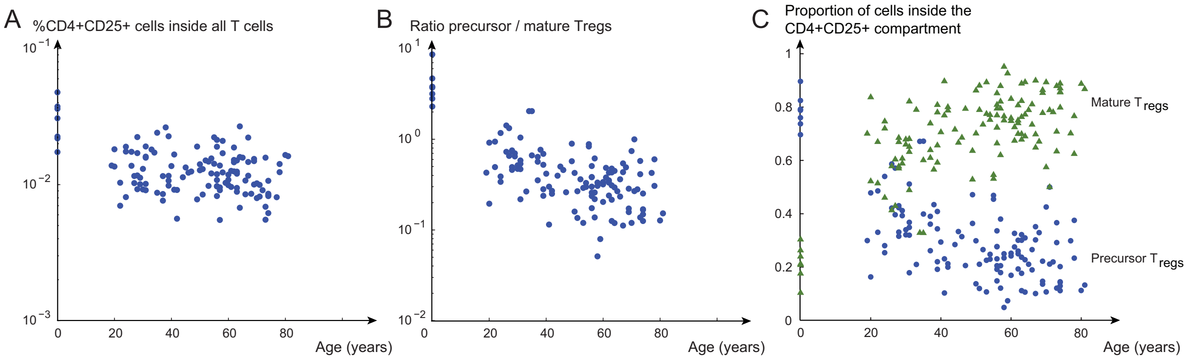


Figure 2.7: Immunological data and properties of T_{regs} dynamics. A. % $CD4^{+}CD25^{+}$ cells inside all T cells; B. Ratio $CD4^{+}CD25^{+}CD45RO^{-}$ precursor / $CD4^{+}CD25^{+}CD45RO^{+}$ mature T_{regs} as a function of age; C. The ratio between precursor and mature T_{regs} is inverted in early adulthood.

Homeostasis is maintained over both $CD4^{+}CD25^{+}CD45RO^{-}$ precursor and $CD4^{+}CD25^{+}$-$CD45RO^{+}$ mature T_{regs} populations

Consider the data of Figure 2.7(A) corresponding to adulthood, i.e., ages 19 to 81. In order to see whether the population of $CD4^{+}CD25^{+}$ cells is in equilibrium, we need to know if there is a significant trend in this data. The best-fit of a linear regression model results in a 95% confidence interval of the slope that contains zero (slope = -0.0034, 95% confidence interval: $[-0.0071, 0.0003]$). Thus the trend is not significant, suggesting that the population of $CD4^{+}CD25^{+}$ T_{regs} is in equilibrium and that there is a common homeostatic mechanism to both $CD4^{+}CD25^{+}CD45RO^{-}$ and $CD4^{+}CD25^{+}CD45RO^{+}$ T_{regs}.

The ratio between $CD4^{+}CD25^{+}CD45RO^{-}$ precursor and $CD4^{+}CD25^{+}CD45RO^{+}$ mature T_{regs} is inverted in early adulthood

By analyzing cord and peripheral blood data for precursor $CD4^{+}CD25^{+}CD45RO^{-}$ and mature $CD4^{+}CD25^{+}CD45RO^{+}$ T_{regs}, we observe that their ratio is inverted before or in early adulthood. Indeed, in Figure 2.7(C), we see from the experimental data giving the proportions of precursor and mature T_{regs} that there is a majority of precursors (~ 80%) in the cord blood of newborns and a majority of mature T$_{\text{regs}}$ (~ 80%) in the peripheral blood of adult donors. This ratio is inverted at latest in the early adulthood (20-30 years of age) and this inversion is what we call the "development of an *in vivo* pool of CD4$^+$CD25$^+$CD45RO$^+$ mature T$_{\text{regs}}$". The following results show that the ratio inversion is a critical issue that is not achieved by all model scenarios.

2.3.2 Analysis of the mean model

In this section, we fit the mean model to the data of Figure 2.7. As a reminder, the mean model is derived from the stochastic one by fixing the parameters of group G_1 to the mean value of their *a priori* distribution and by averaging over the remaining random quantities.

Figure 2.8 shows the trajectories of precursors (left subplots) and mature T$_{\text{regs}}$ (right subplots) under the four proliferation scenarios. Parameters h_P, h_Q, λ, φ_P, φ_Q, Q_M, α, p_{thymus} and s are fitted to the data using a least squares procedure (see Table 4.3 for best-fit values and Table 2.6 for confidence intervals). The three different line types and colors in Figure 2.8 correspond to the three homeostasis scenarios: solid red, scenario (i) (density-independent regulation), dashed blue, scenario (ii) (density-dependent regulation) and dash-dotted green, scenario (iii) (external contribution). The dynamics of precursors are very similar for all scenarios: we observe an increase during the first 10 years, followed by an exponential decrease. This is not surprising as this population is mainly influenced by the influx of recent thymic emigrants. The major difference is in the dynamics of mature T$_{\text{regs}}$. This population increases constantly with age, but according to the studied scenario, we observe qualitative differences in the way the population grows.

Before assessing the proliferation scenarios, we first wanted to see whether one or the other homeostasis scenario is better explaining the data. For this, we applied an F-test for nested models opposing consecutively scenario 1(i) (H_0) to scenarios 1(ii)a, 1(ii)b, 1(iii)a and 1(iii)b (H_1). At level 0.01, H_0 was *not* rejected. This suggests that we do not have a strong evidence that a density-dependent homeostatic mechanism is regulating the population of mature T$_{\text{regs}}$.

The following three subsections describe the assessment results of the four proliferation scenarios, given either homeostasis scenario (i), (ii), or (iii).

In the case of a density-independent regulation of mature T$_{\text{regs}}$, the ratio inversion is a critical issue if the peripheral differentiation of precursors is the only source of mature T$_{\text{regs}}$

First note that due to the linearity of the differential equations of scenario (i), the ratio inversion can be achieved either because of a lifespan difference or through the slow accumulation of mature T$_{\text{regs}}$ following an immune response. In what follows, we'll see that the lifespan difference alone is not sufficient to achieve the ratio inversion.

From the solid red lines of the upper plots in Figure 2.8, we see that scenario 1(i) is achieving the ratio inversion, whereas scenario 2(i) is not. The difference between these two scenarios is easy to explain when observing the typical trajectories of P, Q and R in Figure 2.3. Indeed, the magnitude of immune responses is very different in both scenarios (zoomed sections of Figure 2.3). In scenario 1(i), the clonal expansion goes up to 10^7 cells, whereas in scenario 2(i), it

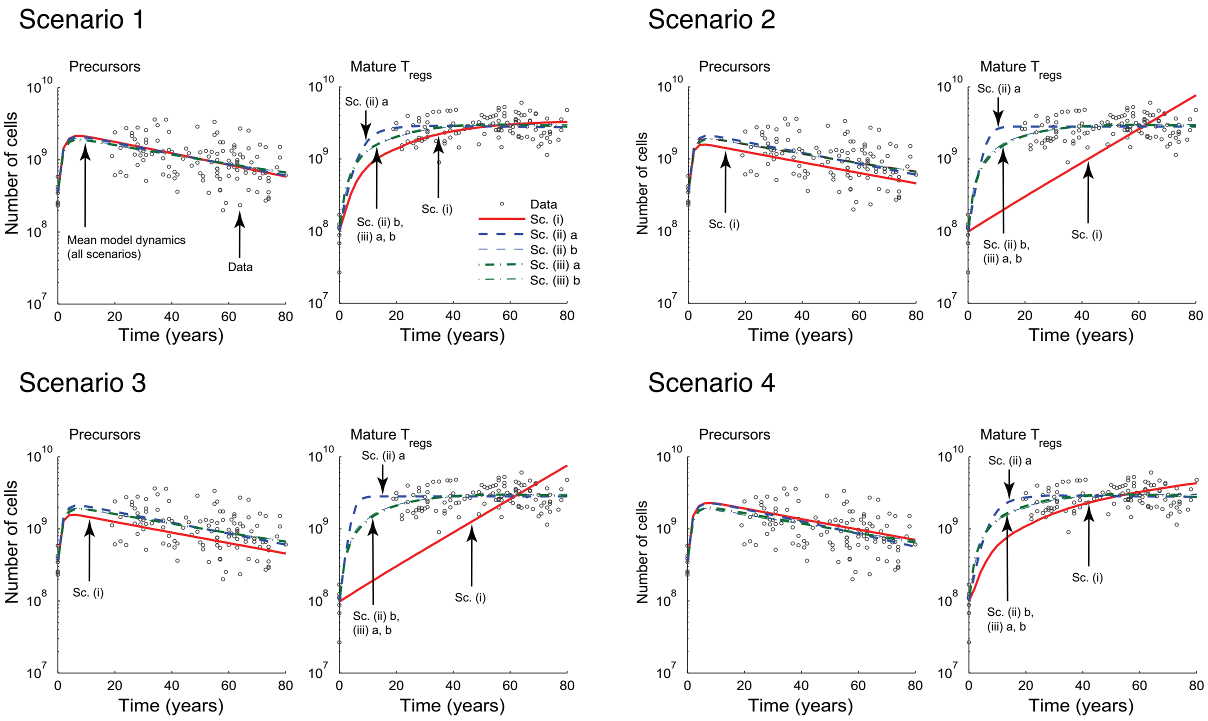


Figure 2.8: Best-fit of the mean model with parameters of group G_1 set to the mean values of their *a priori* distributions. Solid red: for scenario (i) (density-independent regulation); dashed blue: scenario (ii) (density-dependent regulation) and dash-dotted green: scenario (iii) (external contribution). Bold lines: sub-setting (a), light lines: sub-setting (b). The ratio inversion between precursor and mature T$_{\text{regs}}$ is not achieved early enough in scenarios 2(i) and 3(i). All other scenarios are able to explain the data.

does not exceed 10^5 cells. This has an impact on the size of the mature T$_{\text{regs}}$ pool: in the case of scenario 2(i), the only input to the mature T$_{\text{regs}}$ population are the long-lived cells that have survived an immune reaction and they are set as 10% of the maximum effector response. As the peak of the effector response is relatively low, caused by the non-proliferating state of T$_{\text{regs}}$ in scenario 2(i), very few cells feed the mature T$_{\text{regs}}$ pool. Therefore, in scenario 2, the only way to achieve a ratio inversion is through a difference in parameters h_P and h_Q because there is a deficient accumulation of mature T$_{\text{regs}}$ following an antigen response. The latter parameters being fitted, we see that scenario 2(i), even with positive values of h_Q (Table 4.3) can not reach the plateau level of the pool of mature T$_{\text{regs}}$ suggested by the data. This suggests that, in the case of a density-independent regulation of mature T$_{\text{regs}}$, the slow accumulation observed only when there is a sufficient population expansion during antigen-triggered immune reactions is necessary to achieve the ratio inversion.

A similar explanation is valid for scenario 3(i) in which only precursors are endowed with a proliferation capacity. This scenario seems to be unable to accumulate a sufficient pool of mature T$_{\text{regs}}$ (bottom left subplot of Figure 2.8). This is because for the mean values of the parameters of group G_1 (Table 2.4), the differentiation rate b is on average larger than the proliferation rate a_P, meaning that following an antigen priming, precursors have a higher

Scenario		Best-fit Parameters								
		h_P	h_Q	λ	φ_P	φ_Q	Q_M	α	p_{thymus}	s
		($\times 10^{-5}$	($\times 10^{-5}$	(days)	($\times 10^{-13}$	($\times 10^{-13}$	($\times 10^{7}$	(day^{-1})	(%)	($\times 10^{5}$
		day^{-1})	day^{-1})		day^{-1}cell^{-1})	day^{-1}cell^{-1})	cells)			cells)
	(i)	-7.83	-8.87	51.65	4.86					
	(ii) a	-4.42	60.53	58.98	3.27	2.54				
1	(ii) b	7.84	-18.20	99.44	4.41		7.50	0.0051		
	(iii) a	7.10	-8.05	92.94	3.84				83.95	
	(iii) b	14.35	-16.74	95.37	5.11					3.23
	(i)	-0.48	15.16	0.95	6.36					
	(ii) a	-4.69	156.40	60.88	3.23	5.55				
2	(ii) b	12.19	-16.61	164.80	4.87		2.04	0.0250		
	(iii) a	6.54	-1.78	100.95	3.63				78.87	
	(iii) b	12.22	-16.21	97.34	4.87					4.96
	(i)	-0.29	15.07	1.25	6.50					
	(ii) a	-4.75	203.77	55.39	3.22	7.23				
3	(ii) b	12.00	-17.35	168.80	4.85		6.02	0.0090		
	(iii) a	6.55	-1.78	101.74	3.63				78.87	
	(iii) b	12.52	-16.18	98.47	4.90					4.95
	(i)	-20.53	-5.97	63.57	1.48					
	(ii) a	-13.78	64.60	55.70	2.29	2.66				
4	(ii) b	12.15	-19.96	98.57	4.87		9.90	0.0045		
	(iii) a	2.29	-7.96	91.78	3.29				81.87	
	(iii) b	17.46	-16.98	92.24	5.42					3.43

Table 2.5: Best-fit parameters of the mean model dynamics for each model scenario.

probability of differentiating into hyporesponsive mature T$_{\text{regs}}$ than to proliferate as precursors. Thus, most cells become anergic before having divided and the mature T$_{\text{regs}}$ pool is underfed. As in scenario 2(i), a large and positive h_Q (Table 4.3) is not able to make up for the lack of accumulation of mature T$_{\text{regs}}$ following antigen priming. In order to eliminate any dependence of this result to the exact values of the parameters of group G_1, we will study the behavior of scenario 3(i) in the stochastic model (Section 3.3).

Finally, the analysis of the mean model trajectory for scenario 4(i) suggests that the ratio can be inverted (bottom right subplot of Figure 2.8). Note that this scenario is favored by the fact that the rate of secondary infections increases with age. Secondary responses recruit more mature T$_{\text{regs}}$ that, once activated, start proliferating intensively. Therefore, the fact that precursors do not divide has only a minor impact on the population dynamics. However, we will evaluate the performance of this scenario with the stochastic model in order to see how it is affected by a change in the proliferation parameter values.

Scenario		Confidence intervals for the best-fit parameters of the mean model								
		$h_P(\times10^{-5})$	$h_Q(\times10^{-5})$	λ	$\varphi_P(\times10^{-13})$	$\varphi_Q(\times10^{-13})$	$Q_M(\times10^{7})$	$\alpha(\times10^{-2})$	$p_{\text{thymus}}(\%)$	$s(\times10^{5})$
	(i)	[-8.03, -6.73]	[-9.09, -6.91]	[50.75, 65.88]	[2.47, 4.98]					
	(ii) a	[-29.61, 34.81]	[4.32, 230.36]	[46.71, 73.97]	[1.06, 7.37]	[0.57, 8.81]				
1	(ii) b	[-27.99, 66.43]	[-29.76, -2.41]	[82.96, 120.65]	[1.01, 10.51]		[2.00, 11.40]	[0.0, 1.05]		
	(iii) a	[-20.90, 49.82]	[-9.77, -6.38]	[85.39, 100.87]	[1.42, 7.73]				[77.9, 88.9]	
	(iii) b	[-15.47, 72.17]	[-27.86, -11.87]	[77.78, 111.95]	[2.13, 11.12]					[2.11, 6.01]
	(i)	[-0.83, 0.00]	[14.39, 16.66]	[0.79, 93.60]	[2.78, 8.50]					
	(ii) a	[-23.42, 43.02]	[45.72, 273.74]	[14.16, 84.70]	[1.48, 8.33]	[1.53, 9.71]				
2	(ii) b	[-15.26, 72.00]	[-183.19, -11.71]	[79.59, 299.96]	[2.15, 10.64]		[0.94, 415.55]	[0.11, 4.70]		
	(iii) a	[-13.59, 55.69]	[-3.54, -0.23]	[95.15, 119.75]	[1.61, 8.71]				[73.7, 82.9]	
	(iii) b	[-14.51, 71.48]	[-125.20, -10.04]	[52.07, 119.03]	[2.18, 10.86]					[3.42, 34.59]
	(i)	[-0.96, 0.00]	[14.38, 16.44]	[1.00, 120.83]	[2.75, 8.52]					
	(ii) a	[-23.21, 38.08]	[46.20, 278.01]	[5.33, 81.37]	[1.40, 7.49]	[1.53, 10.12]				
3	(ii) b	[-15.48, 72.15]	[-172.93, -11.55]	[80.85, 299.81]	[2.14, 10.67]		[1.00, 407.59]	[0.11, 4.52]		
	(iii) a	[-13.57, 55.92]	[-3.54, -0.22]	[96.77, 119.70]	[1.61, 8.71]				[73.7, 82.9]	
	(iii) b	[-14.49, 71.51]	[-123.99, -10.04]	[49.76, 117.94]	[2.18, 10.86]					[3.42, 34.91]
	(i)	[-28.91, -6.46]	[-8.80, 1.19]	[53.58, 112.02]	[1.11, 2.86]					
	(ii) a	[-28.46, 37.84]	[6.25, 153.84]	[45.90, 70.18]	[1.06, 7.41]	[0.61, 5.73]				
4	(ii) b	[-28.10, 52.84]	[-30.37, -0.62]	[78.67, 121.29]	[1.01, 9.10]		[2.00, 15.89]	[0.01, 1.00]		
	(iii) a	[-18.69, 45.37]	[-9.59, -5.96]	[83.93, 101.44]	[1.43, 7.81]				[76.2, 86.8]	
	(iii) b	[-16.01, 80.73]	[-30.05, -12.50]	[75.17, 118.10]	[2.12, 12.14]					[2.40, 6.68]

Table 2.6: Confidence intervals for the best-fit parameters of the mean model computed using 500 bootstrap replicates.

In presence of a density-dependent homeostatic regulation of mature T$_{\text{regs}}$, the ratio precursor/mature T$_{\text{regs}}$ is inverted without the need of peripheral proliferation in response to antigen.

This claim is suggested by the best-fit of homeostasis scenario (ii), where the population size is regulated either via a density-dependent death (sub-setting (a), bold dashed lines in Figure 2.8), or via a density-dependent homeostatic proliferation (sub-setting (b), light dashed lines in Figure 2.8). We see that both mechanisms produce different dynamics, but the lack of data in the age-range 0-19 years does not allow us to eliminate neither of them. The sum of squared errors indicates however that a density-dependent homeostatic renewal (scenario 1(ii)b) fits slightly better the data. This is not surprising as this scenario has one more parameter than scenario 1(ii)a.

It is interesting to observe that a density-dependent death is compensating for a large and positive value of h_Q (Table 4.3) meaning that the turnover rate of the mature T$_{\text{regs}}$ population is very high. Concerning the density-dependant homeostatic renewal, we see from Table 4.3 that the per cell maximal renewal rate α needs to be on average 2 times larger in the case of hyporesponsiveness to antigen stimulation (scenario 2(ii)b) compared to the case where cells proliferate (scenario 1(ii)b). Therefore we conclude that the density-dependent regulation is able to explain the data as long as there are no restrictions in the parameters defining the homeostatic proliferation and death rates.

The ratio inversion is easy to achieve if there is an exponentially decreasing thymic output or a constant external input to the mature T$_{\text{regs}}$ population

This can be seen by analyzing the performance of homeostasis scenarios (iii)a and b. The green dash-dotted lines in Figure 2.8 show the best-fit of these scenarios to the data (bold: scenario (iii)a, thymic output of mature T$_{\text{regs}}$; light: scenario (iii)b, external source). The ratio precursors/mature T$_{\text{regs}}$ is inverted regardless the underlying proliferation scenario. Although there is no density-dependent regulation here, the increase of the population of mature T$_{\text{regs}}$ is due to the thymic output of mature T$_{\text{regs}}$ (sc. 1-4(iii)a) or to the external and constant input of mature T$_{\text{regs}}$ (sc. 1-4(iii)b) from a phenotype switching of CD4$^+$CD25$^-$ T cells. Thus, we conclude that in this case, the peripheral proliferation of precursor or mature T$_{\text{regs}}$ is not required for the development of the mature T$_{\text{regs}}$ pool.

Furthermore, by analyzing the best-fit parameters of scenarios 1-4(ii)a (Table 4.3), we see that on average, the estimated value of p_{thymus} is about 80%. This means that the data can be explained with 20% of all CD4$^+$CD25$^+$ T cells that exit the thymus having a CD45RO$^+$ mature profile. Assuming that there are 1.3% CD4$^+$CD25$^+$ T$_{\text{regs}}$ inside all CD4$^+$ T cells exiting the thymus, 20% (mature T$_{\text{regs}}$) of all T$_{\text{regs}}$ is equivalent to 0.26% of all CD4$^+$ T cells. Thus, if as few as 0.26% of all CD4$^+$ T cells exiting the thymus are mature T$_{\text{regs}}$, antigen-driven proliferation in periphery is no longer necessary for the deployment of the pool.

2.3.3 Analysis of the stochastic model

The analysis of the mean model led to the rejection of model scenarios 2(i) and 3(i). However, this is true for the mean values of the *a priori* distributions, but it might not be true for all parameter values. To eliminate this possibility, we fit the stochastic model with parameters of group G_1 sampled from the distributions in Table 2.4. We apply this only to homeostasis scenario (i), as the other homeostasis scenarios were able to explain the data already with the mean model.

	a_P (day^{-1})	a_R (day^{-1})	b (day^{-1})
Parameter setting 1	$\mathcal{U}(1,2)$	$a_R = a_P/1.4$	$\mathcal{U}(1,10)$
Parameter setting 2	$\mathcal{U}(1,2)$	$a_R = a_P/1.4$	$\mathcal{U}(0.5,1)$
Parameter setting 3	$\mathcal{U}(1,2)$	$a_R = a_P/5.0$	$\mathcal{U}(0.5,1)$

Table 2.7: Parameter values and distributions of the three parameter settings. U(a, b) stands for a uniform distribution on [a, b].

In order to obtain results that are robust with respect to parameter values, we define three parameter settings (Table 2.7) that give slightly different results. These settings concern three important parameters of the immune response: a_P, the proliferation rate of precursors, a_R, the proliferation rate of activated mature T$_{\text{regs}}$ and b, the differentiation rate of precursors into mature T$_{\text{regs}}$. In the first parameter setting of Table 2.7, the average value of b is larger than the average value of a_P, in other words, precursors have a higher probability of acquiring the mature phenotype than to proliferate as precursors in response to a foreign antigen. In settings 2 and 3, it is the opposite: precursors stay longer (and possibly proliferate) in this undifferentiated state before acquiring the mature phenotype. In settings 1 and 2, a_R takes values 1.4 times smaller than the values of a_P as measured *in vitro*, whereas in setting 3, we model an extreme case where mature T$_{\text{regs}}$ proliferate, but 5 times less than precursors. Note that the analysis of the mean model in Section 3.2 was done under parameter setting 1.

To give an idea of the spread of trajectories of the stochastic model, Figure 2.9 shows the best-fit of the stochastic model under scenarios 1-4(i) (parameter setting 1). In what follows, we use the evaluation procedure described in the Methods to assess all proliferation scenarios and find those that are unable to explain the biological data.

In the absence of a density-dependent homeostatic regulation, and of an external source of CD4$^+$CD25$^+$CD45RO$^+$ mature T$_{\text{regs}}$, proliferation of precursors, or of mature T$_{\text{regs}}$, is necessary to the development of the mature T$_{\text{regs}}$ pool

The result given hereafter is obtained in the case where the unique source of mature T$_{\text{regs}}$ is the peripheral differentiation of precursors without density-dependent homeostatic regulation (scenario (i)). The results of the model evaluation procedure applied to the four proliferation scenarios are shown in Figure 2.10A (see Table 2.8 for the best-fit parameters of the stochastic model). Each bar of Figure 2.10 indicates the absolute value of the maximum likelihood of a proliferation scenario. The closer this bar is to zero, the better the model is explaining the data.

We observe that scenario 2(i), i.e., the one in which neither precursors nor mature T$_{\text{regs}}$ proliferate, has a very poor performance in all three settings of Table 2.7. As stated in the previous section, the problem with this scenario is its inability to achieve the ratio inversion. Indeed, we found no parameter values such that this scenario is explaining the data. Scenario

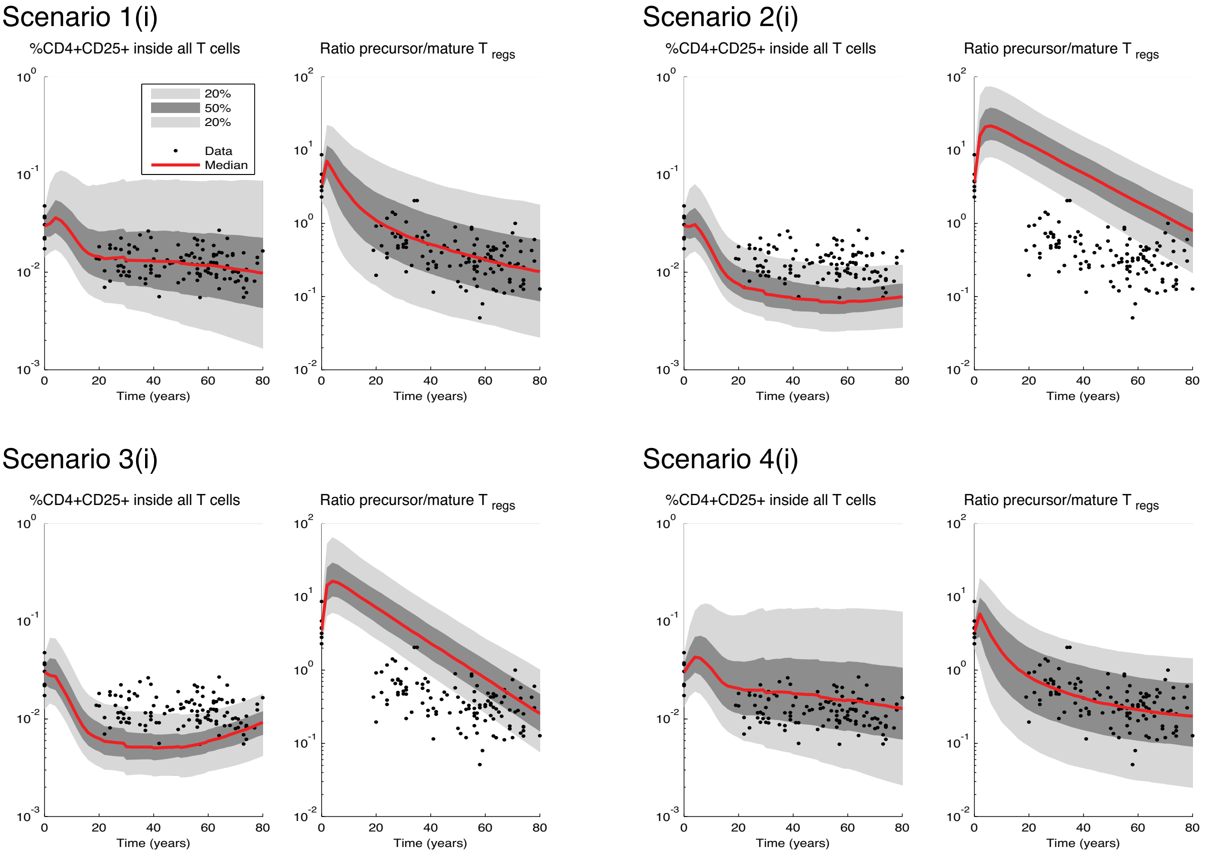


Figure 2.9: Distribution of model's trajectories for scenarios 1(i)-4(i), parameter setting 1, fitted to the data. As the parameters are drawn from probabilistic distributions, every execution of the dynamical system is performed with different parameter values and the trajectories differ. Therefore, instead of a single trajectory, we have a distribution of trajectories. Solid red line: median. 50% of the trajectories are comprised between the 25% and 75% quantiles (dark grey zone); 40% of all trajectories are in the light grey zone, delimited by quantiles 5-25% and 75-95%.

2(i) is therefore rejected when compared to scenario 1(i) using the F-test for nested models (in all parameter settings). This implies that the proliferation of precursors and of mature T_{regs} following a successful interaction with an antigen presenting cell is essential to the development of a stable pool of mature T_{regs} in the absence of a density-dependent homeostatic regulation and of an external source of $CD4^+CD25^+CD45RO^+$ regulatory T cells.

Now, in order to test which proliferating capacity is more important, the one of precursors or the one of mature T_{regs}, we fit the stochastic model scenarios 3(i) and 4(i).

The proliferation of precursors alone is sufficient if $\mathbb{E}[b] < \mathbb{E}[a_P]$

Figure 2.10A shows the results of the evaluation procedure applied to scenario 3(i) (light grey bars) in which only precursors are endowed with a proliferation capacity. We remark that, as suggested by the assessment of the mean model, this scenario performs poorly in parameter

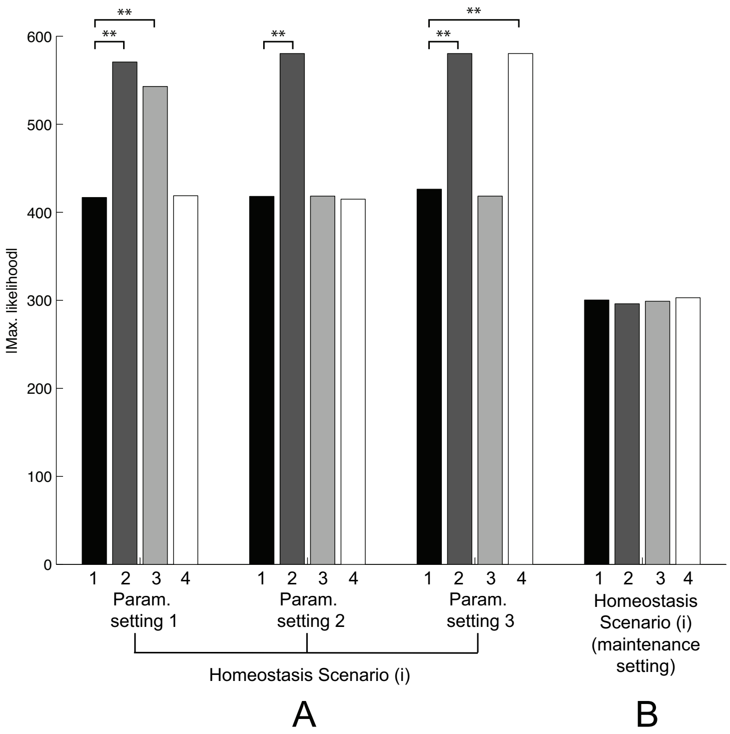


Figure 2.10: Goodness of fit of the mathematical model in scenario (i) (no homeostatic regulation). The bar plot shows the absolute value of the log-likelihood of each model scenario (black bars:proliferation scenario 1, dark grey: prolif. scenario 2, light grey: prolif. scenario 3, white: prolif. scenario 4). The closer the bar is to zero, the better the model fits the data. For each homeostasis scenario, we compare proliferation scenario 1 to all the others by applying a hypothesis test for nested models. A significant difference indicated by ** (p-value < 0.01) means that the considered scenario is rejected.

Homeo. scenario	Prolif. scenario	Parameter setting	Best-fit Parameters			
			$h_P(\times 10^{-5})$	$h_Q(\times 10^{-5})$	λ	$\varphi_P(\times 10^{-13})$
(i)	1	1	-20.00	-5.00	55.22	1.97
		2	-0.12	-17.00	60.90	2.65
		3	-0.18	0.00	6.76	5.72
	2	1	-0.50	8.00	0.67	2.60
		2	-0.50	1.00	1.80	5.30
		3	-0.50	1.00	1.80	5.30
	3	1	-1.00	11.00	5.27	4.78
		2	-1.00	5.00	5.28	4.81
		3	-1.00	5.00	5.28	4.81
	4	1	-9.00	-8.00	20.00	2.00
		2	-0.94	-0.93	22.66	4.45
		3	-0.50	1.00	0.70	4.80
Maintenance (i)	1	1	-0.45	-0.11	200.00	4.07
	2		-0.43	1.00	0.53	4.00
	3		-0.43	1.00	0.53	4.00
	4		-0.57	-1.00	130.00	4.00

Table 2.8: Best-fit parameters of the stochastic model for scenario (i) and all parameter settings.

setting 1. However, using the stochastic model, we have found two other parameter settings (2 and 3) for which scenario 3(i) performs similarly to scenario 1(i). Thus, the stochastic model suggests that there exist parameter values for which the proliferation of precursors alone is sufficient to the development of a viable pool of mature T$_{\text{regs}}$. As a consequence, the F-test for nested models rejects scenario 3(i) only in parameter setting 1. The difference between parameter setting 1 and parameter settings 2 and 3 being in the average value of the differentiation rate b with respect to the value of the proliferation rate of precursors a_P, we conclude that in order to validate model scenario 3(i), it must be shown that the proliferation rate a_P of precursor cells is significantly larger than the rate b of expression of CD45RO following an antigen stimulus.

The proliferation of mature T$_{\text{regs}}$ alone is sufficient if their division rate is not too small compared to the division rate of precursors

The assessment of model scenario 4(i) in which only mature T$_{\text{regs}}$ proliferate is illustrated in Figure 2.10A (white bars). For settings 1 and 2, the performance of this scenario is comparable to the one of scenario 1(i), whereas for parameter setting 3, scenario 4(i) is not able to reproduce the dynamics of the data and the maximum likelihood is fairly low. In that case, the F-test for nested models rejects scenario 4(i) when compared to scenario 1(i). We remark that this happens in the extreme case where mature T$_{\text{regs}}$ proliferate very little (the distribution of a_R in parameter setting 3 is $\mathcal{U}(1/5, 1.7/5)$). Thus, the proliferation of mature T$_{\text{regs}}$ is sufficient for the development of a peripheral pool only if the proliferation rate is not too small.

2.3.4 Maintenance of a lifelong *in vivo* pool of CD4$^+$CD25$^+$CD45RO$^+$ mature T$_{\text{regs}}$

In order to assess the importance of antigen-driven proliferation in the lifelong maintenance of a peripheral pool of T$_{\text{regs}}$, we examine the situation where the model is executed from the time at which the pool is already constituted, i.e., from the age corresponding to our first adulthood data (19 years old). We only study here homeostasis scenario (i), as it is the only one that led to the rejection of some proliferation scenarios.

Given that at the age of 20, a peripheral pool of mature T$_{\text{regs}}$ is constituted, antigen-driven proliferation in the periphery is *not* necessary for the maintenance even in the absence of other input of mature T$_{\text{regs}}$ or of a density-dependent homeostasis mechanism

We assess the four proliferation scenarios of Table 2.3, in the case of homeostasis scenario (i). The percentage of precursors inside CD4$^+$CD25$^+$ T cells at the age of 19 (p_{19}) is drawn from a Gaussian distribution with mean 0.35 and standard deviation 0.1, values obtained from the statistics of the data points corresponding to the age range 19–24 years. The percentage of mature T$_{\text{regs}}$ at the age of 19 is set to $1 - p_{19}$. The maximum likelihood of the fit of the stochastic model under all proliferation scenarios is in Figure 2.10B. None of the scenarios is rejected after the likelihood ratio test. Thus, the proliferation in response to antigen stimuli is not necessary for the lifelong maintenance of a pool of mature T$_{\text{regs}}$, once the pool is constituted.

2.3.5 Summary

All our findings are displayed in Table 2.9.

Biological process		Findings	Found with
Properties of T_{regs} dynamics	1.	Homeostasis is maintained over both $CD4^+CD25^+$-$CD45RO^-$ precursor and $CD4^+CD25^+CD45RO^+$ mature T_{regs} populations.	Data
	2.	The ratio between $CD4^+CD25^+CD45RO^-$ precursor and $CD4^+CD25^+CD45RO^+$ mature T_{regs} is inverted in early adulthood.	Data
Development of the mature T_{regs} pool	3.	The inversion of the ratio precursor/mature T_{regs} is easy to achieve in the presence of a density-dependent homeostatic mechanism of the mature T_{regs} population. In this case, there is no need of proliferation in response to immunogen stimulation.	Model
	4.	The ratio inversion is easy to achieve if there is a thymic or a constant external input to the mature Tregs population.	Model
	5.	In the absence of a density-dependent homeostatic regulation and of an external source of $CD4^+CD25^+CD45RO^+$ mature T_{regs}, proliferation of precursors or of mature T_{regs} is necessary to the development of the mature T_{regs} pool.	Model
	6.	The proliferation of precursors alone is sufficient if the proliferation rate of precursors is significantly larger than the rate of acquisition of CD45RO following an immunogen stimulus.	Model
	7.	The proliferation of mature T_{regs} alone is sufficient if their division rate is not too small compared to the division rate of precursors.	Model
Maintenance of the mature T_{regs} pool	8.	Given that at the age of 20 a peripheral pool of mature T_{regs} is constituted, antigen-driven proliferation in the periphery is not necessary for the maintenance even in the absence of other input of mature T_{regs} or of a density-dependent homeostasis mechanism.	Model

Table 2.9: Summary of all findings.

2.4 Discussion

With the help of a mathematical model, we studied different developmental pathways of regulatory T cells and established conditions in which each of them can explain the biological data. The stochastic infection process combined with the differential equations with random parameters and Monte Carlo simulation is a modeling methodology allowing the description of cell dynamics on the scale of years. We believe that such a scheme is particularly appropriate for use in human studies, where data is sparse and several sources of uncertainty and noise have an influence on parameter values. Although still requiring the computation of confidence intervals, using random parameters was a way of avoiding tedious high-dimensional parameter fitting by decreasing the number of unknown parameters and instead, introducing distributions covering an entire set of possible values. In order to keep the model tractable, we have made several simplifying assumptions. One of them is the assumption that, except for thymic involution, parameter distributions are independent on age. This is probably a simplification, but recent studies indicate that CD4$^+$CD25$^+$ T$_{\text{regs}}$ in aged mice are functionally comparable to those in young mice (Nishioka et al., 2006).

In an effort to keep the model simple, we used a two-phase model (Eq. (2.1) and Eq. (2.3)) that describes the way cells respond to antigen stimulus when only the responder cell population is considered (Althaus et al., 2007; De Boer et al., 2003). Such a model corresponds to a view of the immune system where the size of immune responses is proportional to the number of pre-existing cells. However, the biological reality might be more complicated and it could be possible that the number of cells produced at the end of an immune response is independent of the initial number of cells.

We have observed that the number of CD4$^+$CD25$^+$CD45RO$^-$ precursor T$_{\text{regs}}$ decreases with age. This result confirms the findings of Valmori et al. (2005), where the authors point out a significant negative correlation between the number of precursors and the age of donors. Then, we have observed that the number of CD4$^+$CD25$^+$CD45RO$^+$ mature T$_{\text{regs}}$ increases with age. This finding as well goes in the direction of the results of Valmori et al. (2005) and is in accordance with the results of Gregg et al. (2005), where the authors observe an increased number of peripheral CD4$^+$CD25^{hi} T cells. We then have shown that the ratio between precursors and mature T$_{\text{regs}}$ is inverted in early adulthood. Furthermore, the model suggested the continuous-time trajectory of T$_{\text{regs}}$. In particular, we have observed interesting cell dynamics during infancy and puberty, time-periods for which we have not found immunological data due to ethical reasons. As this is the period of life where the thymus functions at a maximal regime, we have observed an important increase of thymus-derived precursor cells, peaking at 10-12 years old, age corresponding to the onset of puberty. Then, precursor cells start their decline, while mature T$_{\text{regs}}$ continue their progressive increase, having their most important increase rate in the age range 0-20 years. The latter is thus a critical period for the constitution of a pool of mature T$_{\text{regs}}$. This ratio inversion was also a critical issue of the model behavior; it led to the elimination of some model scenarios that could not reproduce it.

We have studied several pool-size regulation mechanisms of T$_{\text{regs}}$: density-independent and density-dependent renewal/death of mature T$_{\text{regs}}$, as well as a constant or time-dependent external contribution. Given these distinct mechanisms, we have identified conditions in which the antigen-driven proliferation is necessary for the creation of a viable pool of CD4$^+$CD25$^+$-

CD45RO$^+$ T$_{\text{regs}}$. It turned out that the antigen-driven proliferation is not necessary if a density-dependent homeostatic mechanism regulates the pool of mature T$_{\text{regs}}$. In other terms, as we found no information about the exact values of the parameters governing the homeostatic regulation, the antigen-driven cell division could be replaced in our model by a homeostatic (cytokine-driven) proliferation. However, the homeostatic expansion is unlikely to be the main source of mature T$_{\text{regs}}$, because if that was the case, the diversity of the T cell receptor (TCR) repertoire would be reduced by competitive exclusion. Indeed, if there is no external contribution to this cell population, the diversity of T cell receptors would be determined by the few cells present at birth. Yet, it is known that T$_{\text{regs}}$ have an $\alpha\beta$ TCR repertoire with size and diversity closely similar to those of CD4+CD25- T cells (Fazilleau et al., 2007). Therefore, we believe that in addition to a homeostatic mechanism, other sources of *de novo* generation should be present.

In that perspective, we have studied other sources of mature T$_{\text{regs}}$ that are able to replenish the peripheral pool of mature CD4$^+$CD25$^+$CD45RO$^+$ T$_{\text{regs}}$ and enrich the TCR repertoire. One such source is the thymus output of mature T$_{\text{regs}}$. We have established the percentage of T$_{\text{regs}}$ that should be output daily from the thymus so that the peripheral expansion of this population is not required for a successful development of the pool. This threshold is about 20% of all CD4$^+$CD25$^+$ T cells output from the thymus, or, equivalently, less than 0.3% of all CD4$^+$ T cells exiting the thymus. Another question arises then: are T$_{\text{regs}}$ with a mature phenotype able to exit the thymus? We found two points of view in the literature.

On the one hand, Wing et al. (Wing et al., 2003) claim that only regulatory T cells with a naive phenotype, namely CD45RO- precursors, are released in the periphery from the thymus (corresponding to our homeostasis scenario (i)). If this is indeed the case, then we have demonstrated that proliferation of both CD4$^+$CD25$^+$CD45RO$^-$ and CD4$^+$CD25$^+$CD45RO$^+$ cells is sufficient, but not necessary for the development and maintenance of a pool of T$_{\text{regs}}$. In scenario 3(i), where only precursors proliferate, we saw that if the differentiation rate of CD4$^+$-CD25$^+$CD45RO$^-$ precursors into CD4$^+$CD25$^+$CD45RO$^+$ mature T$_{\text{regs}}$ is similar to or greater than the proliferation rate of precursors, the model cannot explain the data. However, it is unlikely that precursor cells stop suddenly their proliferation because of the acquisition of the mature phenotype. Our *in vitro* experiments have shown that in presence of antigen with a co-stimulatory signal, precursor CD4$^+$CD25$^+$CD45RO$^-$ cells proliferate intensively while acquiring the memory phenotype (unpublished data). This observation suggests that either the acquisition of CD45RO is not associated with anergy during a first priming, in which case, instead of scenario 3(i), scenario 1(i) would be the appropriate model, or the parameters governing cell dynamics have the property that the rate of proliferation of precursors is greater than the rate of differentiation into mature T$_{\text{regs}}$, in which case scenario 3(i) is in accordance with the *in vivo* data. In scenario 4(i), where only activated recently-matured CD4$^+$CD25$^+$-CD45RO$^+$ cells proliferate, cells start their intense proliferation after a time-delay, which is compatible with the fact that regulatory T cells activate slowly, only when there is a sufficient amount of growth resources in the environment. The fact that the expansion of regulatory T cells is dependent on growth factors produced by other cells was also pointed out in (Burroughs et al., 2006; León et al., 2000) . A recent study (Lee et al., 2008) brings light to the molecular mechanisms through which Foxp3 maintains the unresponsiveness of T$_{\text{regs}}$. The link of Foxp3 to c-Jun blocks proliferation and therefore makes it even more difficult for T$_{\text{regs}}$ to activate and proliferate. Scenario 4 with its initial delay, attests for this difficulty.

On the other hand, it cannot be excluded that a small proportion of mature-type T$_{\text{regs}}$ might exit thymus tissues. Vanhecke et al. (1995) identify five stages of thymocyte differentiation during which CD4$^+$ T cells acquire and lose several surface receptors. The last two stages define the most mature cells that consist of CD4$^+$CD1$^-$CD45RO$^+$ (stage 4) and CD45RO$^-$ (stage 5) helper T cells. As no particular sorting was applied to mark regulatory T cells in this study, we can assume that they are comprised in the above populations. Vanhecke et al. (1995) claim that both stage 4 and stage 5 cells emigrate from the thymus in a severe combined immunodeficient mouse carrying a human thymus, even though they consider CD45RO- stage 5 cells as more mature. Another question arises then, which is whether these thymic emigrants with mature profile switch back to a naive form soon after their entry in the periphery or not. This event was not considered in our mathematical model, but it will certainly reinforce the need of proliferating capacity of T$_{\text{regs}}$ for the development of a reliable pool. However, if these stage 4 CD45RO$^+$ T$_{\text{regs}}$ maintain the mature profile once in the periphery, their number could be sufficient for the development of a pool of mature T$_{\text{regs}}$.

Recent studies point out the presence of an input of mature T$_{\text{regs}}$ coming from non-regulatory helper T cells (Vukmanovic-Stejic et al., 2006). This hypothesis is perfectly able to explain the biological data, when daily there are at least 10^5 CD4$^+$CD25$^-$ cells acquiring a regulatory profile. To our knowledge, there is no quantitative data measuring this transformation, nonetheless, we hypothesize that this external contribution should be such that the entire pool of CD4$^+$CD25$^+$FoxP3$^+$ cells, including naturally occurring thymus-derived T$_{\text{regs}}$, is in steady state, as suggested by our data.

At last, we have shown that although the antigen-driven proliferation of regulatory T cells is essential for the development of a pool of mature T$_{\text{regs}}$, it is not a critical issue for the lifelong maintenance of this compartment. This finding confirms the intuitive fact that the homeostasis-related kinetics have more impact on the lifelong maintenance of a cell population than the antigen-response kinetics.

An important contribution of the mathematical model described here is the suggestion of new research directions that will deepen our knowledge about regulatory T cells. Indeed, our results led to the identification of crucial mechanisms having an important impact on T$_{\text{regs}}$ dynamics and that could be subject to further investigations: the quantitative assessment of the parameters governing the homeostatic maintenance of T$_{\text{regs}}$, the assessment of the quantity of mature T$_{\text{regs}}$ exiting the thymus, and the quantitative description of the phenomena transforming a non-regulatory cell into a regulatory CD4$^+$CD25$^+$FoxP3$^+$ T cell *in vivo*.

In conclusion, the mathematical model allowed us to have a global view on the lifelong dynamics of T$_{\text{regs}}$. We evaluated three different homeostatic scenarios about possible sources of mature T$_{\text{regs}}$ that can explain how a stable pool of T$_{\text{regs}}$ is developed and maintained through human life. This enabled us to appreciate and estimate the different contributions of each single path in the absence of others. Moreover, we also evaluated four different scenarios concerning the intrinsic proliferation ability of precursors and mature T$_{\text{regs}}$. This ability has a considerable impact on T$_{\text{regs}}$ development and maintenance if precursors are the unique source of mature T$_{\text{regs}}$. Our model is the first attempt to mathematically and computationally describe the behavior of regulatory T cells over life. The mathematical model is able to estimate the trend of T$_{\text{regs}}$ over time when one or more sources are affected. Although it is already possible to have an empirical idea of the future trend of regulatory T cells from the blood of a patient, it is not possible to predict their amount at a certain time in the future. Further developments could allow our *in silico* model to virtually monitor and predict the dynamics of regulatory T cells through an individual life. This application can be particularly useful for prospective studies on patients that received a solid organ transplant or are suffering from an autoimmune disease. Such patients have often decreased numbers of $\mathrm{T_{regs}}$ in their blood when compared to healthy donors. Once the mathematical model is able to predict the amount of $\mathrm{T_{regs}}$ over time, it would allow to individually dose the immunosuppressive drugs needed to prevent chronic rejection episodes or disease relapses. In the perspective of introduction of regulatory T cells in cell-therapy as a more specific and sophisticated substitute to immunosuppressive drugs (Riley et al., 2009), our model is a first step towards the development of tools aiming the clinical monitoring of regulatory T cell dynamics before and after their adoptive transfer.

Chapter 3

Analysis of the T$_{\text{regs}}$ Model

In this chapter are given theoretical results related to the model of regulatory T cells presented in Chapter 2. In Section 3.1, we obtain an analytical expression of the dynamical system in question by solving the differential equations Eq. (2.1), Eq. (2.3) and Eq. (2.3) in the particular case where these are linear (in the absence of homeostasis). In the following section (Section 3.2), we present the detailed computation of the mean dynamics, again in the absence of homeostasis. Finally, in Section 3.3, we describe how is estimated the density of model trajectories, needed to compute the maximum likelihood of a scenario.

3.1 Generic model solution

In this section, we consider the T$_{\text{regs}}$ model under homeostasis scenario (i), i.e., where there is no competition between cells for a common resource, nor fratricide. In this case, $g_Q(t) = 0, s_Q = 0, \varphi_Q = 0, \alpha = 0$ and the ODE system of Eq. (2.3) is linear. As the ODEs describing immune responses are linear as well, we can easily solve the generic model. To do that, we write the differential equations corresponding to the n^{th} immune reaction at time τ_n in matrix form:

$$\frac{dY_{\text{clone}}(t)}{dt} = \begin{cases} A\, Y_{\text{clone}}(t) & \text{if } \tau_n \leq t \leq \delta \\ B\, Y_{\text{clone}}(t) & \text{if } \delta \leq t \leq \tau_{n+1} \end{cases}, \tag{3.1}$$

where

$$A = \begin{pmatrix} a_P - b & 0 & 0 \\ 0 & -f & 0 \\ b & f & a_R \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} -d_P & 0 & 0 \\ 0 & -d_Q & c \\ 0 & 0 & -(c + d_R) \end{pmatrix}$$

are respectively the matrices defining the expansion and contraction phases of the immune reaction. Solving Eq. (3.1) for $t \in [\tau_n, \tau_{n+1}]$ yields:

$$Y_{\text{clone}}(t) = e^{Bt} e^{-B(\tau_n + \delta)}\, e^{A\delta}\, Y_{\text{clone}}(\tau_n), \tag{3.2}$$

where the matrix exponential of a square matrix X is defined as $e^X = \sum_{k=0}^{\infty} \frac{1}{k!} X^k$.

The homeostatic activity of the cells $Y_{\text{out}}(\tau_n)$ that are not part of a clone at the time of occurrence of the n^{th} immune reaction can be written as:

$$\frac{dY_{\text{out}}(t)}{dt} = f(t) + CY_{\text{out}}(t), \tag{3.3}$$

where

$$f(t) = \begin{pmatrix} g_P(t) \\ g_Q(t) + s_Q \\ 0 \end{pmatrix} \quad \text{and} \quad C = \begin{pmatrix} -h_P & 0 & 0 \\ 0 & -h_Q & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

represent respectively the thymic output and the homeostasis matrices. Solving Eq. (3.3) for $t \in [\tau_n, \tau_{n+1}]$ yields:

$$Y_{\text{out}}(t) = F(t) + e^{C(t-\tau_n)}\, Y_{\text{out}}(\tau_n), \tag{3.4}$$

where

$$F(t) = \begin{pmatrix} \frac{\sigma_P}{h_P-\nu}\left(e^{-\nu t} - e^{-\nu\tau_n - h_P(t-\tau_n)}\right) \\ \frac{\sigma_Q}{h_Q-\nu}\left(e^{-\nu t} - e^{-\nu\tau_n - h_Q(t-\tau_n)}\right) + \frac{s_Q}{h_Q}\left(1 - e^{-h_Q(t-\tau_n)}\right) \\ 0 \end{pmatrix}.$$

We obtain the expression defining the generic model dynamics between two consecutive immune reactions by evaluating Eq. (3.2) and Eq. (3.4) at time τ_{n+1} and by using the fact that $Y_{\text{out}}(\tau_n) = Y(\tau_n) - Y_{\text{clone}}(\tau_n)$:

$$Y(\tau_{n+1}) = \left(e^{B(\Delta_n-\delta)}\, e^{A\delta} - e^{C\Delta_n}\right) Y_{\text{clone}}(\tau_n) + e^{C\Delta_n} Y(\tau_n) + F(\tau_n + \Delta_n), \tag{3.5}$$

where $\Delta_n = \tau_{n+1} - \tau_n$. The probability that the n^{th} immune reaction is primary (resp. secondary) is given by:

$$\begin{aligned} \mathbb{P}(Y_{\text{clone}}(\tau_n) = \pi_n(P(\tau_n),\ 0,\ 0)') &= q(\tau_n) \\ \mathbb{P}(Y_{\text{clone}}(\tau_n) = \pi_n(P(\tau_n),\ Q(\tau_n),\ 0)') &= 1 - q(\tau_n). \end{aligned}$$

The initial condition (at birth: $t = 0$) is $Y(0) = (P_0, Q_0, 0)'$.

3.2 Mean model dynamics

Here, we consider again the simple case of homeostasis scenario (i), and we study the model dynamics when all parameter values are set to their mean value. For this purpose, we assume in this section that all parameters are fixed to their mean value, but to keep the notation simple, we use in what follows the same symbols as previously.

The mean system dynamics are computed from Eq. (3.5) by taking the expectation over the stochastic quantities of the model. Let $\xi_n = (Y_{\text{clone}}(\tau_n), \Delta_n, \pi_n)$ be the vector of random variables in our model. The distribution of $Y_{\text{clone}}(\tau_n)$ is Bernoulli with parameter q and π_n is uniform over the range $[10^{-7}, 10^{-4}]$, $\forall n$. The distribution of Δ_n is the exponential distribution shifted by the constant δ and its density is given by

$$\phi_{\Delta_n}(x) = \begin{cases} \frac{1}{\lambda} e^{-\frac{(x-\delta)}{\lambda}} & \text{if} \quad x > \delta \\ 0 & \text{otherwise.} \end{cases}$$

We can write:

$$
\begin{aligned}
\mathbb{E}_{\xi_n}[Y(\tau_{n+1})] &= \mathbb{E}_{\xi_n}\left[\left(e^{B(\Delta_n-\delta)}\, e^{A\delta} - e^{C\Delta_n}\right)\, Y_{\text{clone}}(\tau_n)\right] \\
&\quad + \mathbb{E}_{\xi_n}\left[e^{C\Delta_n}\right] Y(\tau_n) + \mathbb{E}_{\xi_n}[F(\tau_n+\Delta_n)] \\
&= \mathbb{E}_{\Delta_n}\left[\left(e^{B(\Delta_n-\delta)}\, e^{A\delta} - e^{C\Delta_n}\right)\right]\, \mathbb{E}_{Y_{\text{clone}}(\tau_n),\pi_n}[Y_{\text{clone}}(\tau_n)] \\
&\quad + \mathbb{E}_{\xi_n}\left[e^{C\Delta_n}\right] Y(\tau_n) + \mathbb{E}_{\Delta_n}[F(\tau_n+\Delta_n)],
\end{aligned}
$$

where we have used the fact that $Y(\tau_n)$ is independent of ξ_n and that $Y_{\text{clone}}(\tau_n)$ is independent of Δ_n. If I is the identity matrix, we have the following results:

$$
\begin{aligned}
\mathbb{E}_{\Delta_n}\left[e^{B(\Delta_n-\delta)}\right] &= (I-\lambda B)^{-1} \\
\mathbb{E}_{\Delta_n}\left[e^{C\Delta_n}\right] &= e^{C\delta}(I-\lambda C)^{-1} \\
\mathbb{E}[Y_{\text{clone}}(\tau_n,\pi_n)] &= q(\tau_n)\, \mathbb{E}_{\pi_n}[\pi_n(P(\tau_n),0,0)'] \\
&\quad +(1-q(\tau_n))\, \mathbb{E}_{\pi_n}[\pi_n(P(\tau_n),Q(\tau_n),0)'] \\
&= \begin{pmatrix} \bar{\pi} & 0 & 0 \\ 0 & (1-q(\tau_n))\bar{\pi} & 0 \\ 0 & 0 & 0 \end{pmatrix} Y(\tau_n) = U(\tau_n)\, Y(\tau_n) \qquad (3.6) \\
\mathbb{E}_{\Delta_n}[F(\tau_n+\Delta_n)] &= \lambda \begin{pmatrix} \frac{\sigma_P e^{-\nu\tau_n}}{(1+\lambda\nu)(1+\lambda h_P)} \\ \frac{\sigma_Q e^{-\nu\tau_n}}{(1+\lambda\nu)(1+\lambda h_Q)} + \frac{s_Q}{1+\lambda h_Q} \\ 0 \end{pmatrix} = F^*(\tau_n). \qquad (3.7)
\end{aligned}
$$

If we let $Y^*(\tau_n) = \mathbb{E}_{\xi_n}[Y(\tau_n)]$, we can write

$$
\begin{aligned}
Y^*(\tau_{n+1}) &= \underbrace{\left[\underbrace{\left[(I-\lambda B)^{-1}e^{A\delta} - (I-\lambda C)^{-1}e^{C\delta}\right]U(\tau_n)}_{\text{infection matrix: } M_I} + \underbrace{(I-\lambda C)^{-1}e^{C\delta}}_{\text{pool-size control: } M_H}\right]}_{J^* = M_I(\tau_n)+M_H} Y^*(\tau_n) + \underbrace{F^*(\tau_n)}_{\text{thymus}} \\
&= J^*(\tau_n)Y^*(\tau_n) + F^*(\tau_n), \qquad (3.8)
\end{aligned}
$$

where $U(\tau_n)$ is given by the square matrix in Eq. (3.6) and $F(\tau_n)$ is given by Eq. (3.7). The mean model dynamics are obtained by applying the iteration of Eq. (3.8) to the initial condition $Y^*(0) = (\bar{P}_0, \bar{Q}_0, 0)'$, where $\bar{P}_0$ and $\bar{Q}_0$ are the average values of P_0 and Q_0.

3.3 Model density estimation

In order to estimate the model density accurately, we computed confidence intervals on the bivariate histogram depicted in Figure 2.6A of Chapter 2. In this section, we describe how this is done. This is important because this density is used to formally reject a model scenario.

First, we introduce the necessary theoretical framework. We have a data set $X = (x_1, \ldots, x_n)$ that we view as the realization of a stochastic system (the output of a simulator). We usually assume that the model has a density of probability, and that the output $(x_1, \ldots, x_n)$ depends on a parameter θ; we denote this density by $f(x_1,\dots,x_n|\theta)$. It is also called the *likelihood* of the observed data. The framework of parametric estimation theory consists in assuming that the parameter θ is fixed, but unknown. An estimator of θ is any function $T(\cdot)$ of the observed data. The maximum likelihood estimator (MLE) is the value of θ that maximizes the likelihood $f(x_1,\dots,x_n|\theta)$; we denote it by $\hat{\theta}$. This definition makes sense if the maximum exists and is unique, which is often true in practice. Definition 1 gives a formal set of conditions that guarantee the existence and uniqueness of the MLE.

Assume that the parameter θ is multidimensional, i.e., it varies in an open subset Θ of $\mathbb{R}^k$. The *observed information* is defined by the symmetric matrix

$$[J(\theta)]_{i,j} = -\frac{\partial^2 \ell(\theta)}{\partial\theta_i \partial\theta_j},$$

where $\ell(\theta)$ is the log-likelihood defined by

$$\ell(\theta) = \log(f(x_1,\dots,x_n|\theta)).$$

The *Fisher information*, or *expected information* is defined by the matrix

$$[I(\theta)]_{i,j} = -\mathbb{E}_\theta(J(\theta)) = -\mathbb{E}_\theta\left(\frac{\partial^2 \ell(\theta)}{\partial\theta_i \partial\theta_j}\right),$$

where the notation $\mathbb{E}_\theta$ means that the expectation is a function of θ.

Before stating the asymptotic result that leads to the confidence intervals of θ, we state the regularity conditions.

Definition 1. *Regularity conditions for maximum likelihood asymptotics.*

1. *The set Θ of values of θ is compact (closed and bounded) and the true value $\hat{\theta}$ is not on the boundary.*
2. *For different values of θ, the densities $f(X|\theta)$ are different.*
3. *There exists a neighborhood B of θ^* and a constant K such that for $\theta \in B$ and for all i, j, k, n, we have $\frac{1}{n}\mathbb{E}_\theta(|\partial^3 \ell_X(\theta)/\partial\theta_i\partial\theta_j\partial\theta_k|) \leq K$.*
4. *For $\theta \in B$, the Fisher information has full rank.*
5. *For $\theta \in B$, the interchanges of integration and derivation in $\int \frac{\partial f(X|\theta)}{\partial\theta_i}\,dx = \frac{\partial}{\partial\theta_i}\int f(X|\theta)\,dx$ and in $\int \frac{\partial^2 f(X|\theta)}{\partial\theta_i\partial\theta_j}\,dx = \frac{\partial}{\partial\theta_i}\int \frac{\partial f(X|\theta)}{\partial\theta_j}\,dx$ are valid.*

The following theorem is proven in Davison (2003).

Theorem 1. *Under the conditions of Definition 1, the MLE $\hat{\theta}$ exists and converges almost surely to the true value. Further, $I(\theta)^{1/2}(\hat{\theta}-\theta)$ converges in distribution to a standard normal distribution, as n goes to infinity. It follows that, asymptotically:*

1. *the distribution of $\hat{\theta}-\theta$ can be approximated by $\mathcal{N}(0, I(\hat{\theta})^{-1})$ or $\mathcal{N}(0, J(\hat{\theta})^{-1})$;*

2. *the ditribution of* $2(\ell(\hat{\theta}) - \ell(\theta))$ *(also called the likelihood ratio statistic) can be approximated by a chi-square distribution with* k *degrees of freedom, where* k *is the dimension of* Θ.

We use the following corollary to define the confidence interval on θ.

Corollary 1. *(Asymptotic confidence intervals) When* n *is large, an approximate confidence interval for the* i^{th} *coordinate of* θ *is*

$$\hat{\theta}_i \pm \eta\sqrt{\left[I(\hat{\theta})^{-1}\right]_{i,i}} \quad or \quad \hat{\theta}_i \pm \eta\sqrt{\left[J(\hat{\theta})^{-1}\right]_{i,i}}, \tag{3.9}$$

where $\mathcal{N}_{0,1}(\eta) = \frac{1+\gamma}{2}$ *(for example, with* $\gamma = 0.95$, $\eta = 1.96$*).*

We apply the above corollary to the estimation of our model's density. The statistic that we use is the bivariate histogram of U and V. Remind that $U(t) = \log(P(t)/(Q(t) + R(t)))$ and $V(t) = \log((P(t) + Q(t) + R(t))/T(t))$, where $T(t)$ is the total number of T cells at time t. To simplify the notation, consider for the moment a univariate histogram, built from n independent observations that fall into categories $1, \dots, k$ (the bivariate case is straightforward). Let Y_i denote the number of observations in category i. If π_i is the probability that a single observation falls into category i ($0 < \pi_i < 1$ and $\sum_{i=1}^{k} \pi_i = 1$), the random variable $\vec{Y} = (Y_1, \dots, Y_k)'$ is multinominal with probabilities $(\pi_1, \dots, \pi_k)$ and denominator n. We want to estimate the parameters $\theta = (\pi_1, \dots, \pi_k)'$ of the multinomial. As $\pi_k = 1 - \pi_1 - \cdots - \pi_{k-1}$, the parameter space is the interior of a simplex in k dimensions, that is, the set

$$\left\{(\pi_1, \dots, \pi_k) : \sum_{i=1}^{k} \pi_i = 1, 0 < \pi_1, \dots, \pi_k < 1\right\}$$

of dimension $k - 1$. Therefore, there are $k - 1$ parameters to estimate.

Given data $y_1, \dots, y_k$, the likelihood under the multinomial model is

$$L(\theta) = \frac{n!}{y_1! \cdots y_k!} \pi_1^{y_1} \times \cdots \times \pi_k^{y_k}, \quad \sum_{i=1}^{k} \pi_i = 1, \quad 0 < \pi_1, \dots, \pi_k < 1,$$

where $\sum_i y_i = n$, so the log-likelihood is

$$\ell(\theta) \simeq \sum_{i=1}^{k-1} y_i \log \pi_i \; + \; y_k \; \log(1 - \pi_1 - \cdots - \pi_{k-1}).$$

The observed information matrix has general term

$$-\frac{\partial^2 \ell(\theta)}{\partial \pi_i \partial \pi_j} = \begin{cases} \frac{y_i}{\pi_i^2} + \frac{y_k}{(1-\pi_1-\cdots-\pi_{k-1})^2}, & \text{for } i = j \\ \frac{y_k}{(1-\pi_1-\cdots-\pi_{k-1})^2} & \text{for } , i \neq j, \end{cases},$$

where i and j run over $1, \dots, k-1$. It is easy to see that the maximum likelihood estimators are $\hat{\pi}_i = Y_i/n$. The expected information matrix involves $\mathbb{E}(Y_i)$, which is calculated by noting that the marginal distribution of Y_i is binomial with probability π_i and denominator n. The mean of Y_i is therefore $n\pi_i$. Thus, the expected information matrix is:

$$I(\theta) = n\begin{pmatrix} 1/\pi_1 + 1/\pi_k & 1/\pi_k & \cdots & 1/\pi_k \\ 1/\pi_k & 1/\pi_2 + 1/\pi_k & \cdots & 1/\pi_k \\ \vdots & \vdots & \ddots & \vdots \\ 1/\pi_k & 1/\pi_k & \cdots & 1/\pi_{k-1} + 1/\pi_k \end{pmatrix}$$

whose inverse is:

$$I(\theta)^{-1} = \frac{1}{n}\begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & \cdots & -\pi_1\pi_{k-1} \\ -\pi_2\pi_1 & \pi_2(1-\pi_2) & \cdots & -\pi_2\pi_{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ -\pi_{k-1}\pi_1 & -\pi_{k-1}\pi_2 & \cdots & \pi_{k-1}(1-\pi_{k-1}) \end{pmatrix}. \tag{3.10}$$

Provided that none of the π_i equals zero or one, the regularity conditions of Definition 1 hold and Theorem 1 applies with $I(\theta)^{-1}$ given by Eq. (3.10), giving us the confidence interval of $\hat{\pi}_i$:

$$\hat{\pi}_i \pm \eta\sqrt{\left(\frac{1}{n}\pi_i(1-\pi_i)\right)}, \tag{3.11}$$

where $\mathcal{N}_{0,1}(\eta) = \frac{1+\gamma}{2}$.

Part II

Measuring T-cell Receptor Diversity

Chapter 4

Mathematical Modeling of AmpliCot

4.1 Introduction

The diversity of T cell receptors (TCRs) is part of the polymorphism of our adaptive immune system and is responsible for the recognition and the defense against possibly an universe of different pathogens. The structural diversity of TCRs is achieved by somatic gene-segment rearrangements and random nucleotide additions or deletions (Goldsby et al., 2003). The estimation of the effective size of the human TCR repertoire, both in health and disease, is a fundamental question in immunology and yet, not fully addressed.

A common way to tackle this problem consists in the following two steps: (1) sampling the repertoire of an individual to measure the sample diversity; (2) extrapolating the whole repertoire diversity from the sample diversity. The first step of this *"general approach"* is often rather experimental, whereas the second is rather theoretical, as discussed in Section 4.1.1. The work presented in this chapter concerns the first step above and combines experiments and mathematical modeling in order to better estimate the diversity of a sample.

Several experimental techniques aim at the measurement of the structural TCR diversity of a repertoire sample. Immunoscope (or Spectratype) gives a qualitative insight of the repertoire's shape in terms clonal sizes (Currier and Robinson, 2001; Pannetier et al., 1993); high-throughput DNA sequencing exhaustively enumerates the clonotypes of a sample, thus providing a more detailed picture of the repertoire (Mardis, 2008; Shendure and Ji, 2008). AmpliCot is an alternative experimental technique that allows the sample diversity measurement through quantitation of the re-hybridization speed of denatured PCR products (Baum and McCune, 2006). This elegant approach has the advantage over the cloning and sequencing methods to be time- and expense- effective. However, in order to obtain accurate diversity estimates, the assay should be performed under very stringent experimental conditions.

The AmpliCot experiment is based on the so-called "Cot analysis" (Britten and Kohne, 1968), according to which the time required for a DNA sample to reanneal (expressed in terms of the product nucleotide concentration × time, "Cot") is related to the diversity of the sample. In short, fluorescent SYBR green dye, which binds double-stranded DNA, is added to a sample of PCR-amplified cDNA. The sample is melted, such that the DNA becomes single-stranded, and reannealed under very strict conditions in order to avoid heteroduplex formation and allow only for the association of perfectly complementary strands (homoduplexes).[1] The annealing kinetics, measured in terms of fluorescence intensity, are a function of the diversity and of the concentration of the sample, and have been interpreted assuming second order kinetics.

In order to read the diversity from the resulting annealing curve, Baum and McCune (2006) propose a simple method. It consists in considering the Cot value at which 50% of the sample is annealed ($\mathrm{Cot}_{1/2}$ value) as a function of diversity. The authors suggest that the relation between these quantities is linear. This statement presumes the validity of *second order kinetics* (SOK) as a correct model for the annealing kinetics (Yguerabide and Ceballos, 1995). Accordingly, only perfectly complementary pairs of DNA can associate, precluding the possibility of heteroduplex formation. Under the “stringent" conditions defined in Baum and McCune (2006), SOK would indeed hold. However, it is possible that a slight deviation from the “ideal" experimental settings leads to some heteroduplex formation, which would bias the diversity estimation.

We therefore have two motivations to define a mathematical model of AmpliCot. First, driven by the fact that SOK might be an over-simplified model, we define a more detailed scheme in which the formation of transient complexes and heterdoduplexes is allowed. Second, we consider that sampling the annealing curve at a single point (the $\mathrm{Cot}_{1/2}$ value) is not taking advantage of all the available information in the data. We believe that the diversity extrapolation method can be improved by using the information contained in the entire annealing curve.

Thus, in this chapter, we define a mathematical model that describes the annealing kinetics of AmpliCot and use it to propose an alternative diversity extrapolation procedure. To fit the model parameters and to test our diversity predictions, we use “toy" data sets consisting in templates of known diversity from a library of individually synthesized oligonucleotides.

4.1.1 TCR repertoire assessment: State of the art

In this section, we give a (non-exhaustive) overview of the recent advances in the TCR repertoire assessment. In particular, we focus on studies using mathematical modeling or advanced statistical analysis. Although the experimental settings evoked here can be used to measure the structural diversity of BCRs and antibodies, we consider only examples in which they have been used to study TCRs. In addition, a large part of the existing diversity assessment literature concerns the repertoire of $\alpha\beta$ T cells. Therefore, unless explicitly stated, TCR diversity refers in this section to the structural diversity of $\alpha\beta$ TCRs.

The main difficulty in the estimation of TCR diversity comes from the fact that the latter has to be extrapolated from samples containing only a small part of the entire repertoire. The difficulty is further increased by the unknown shape of the clonal sizes distribution. A lower bound estimate was obtained by assuming that all clonotypes are equally represented, as in Arstila et al. (1999). Using TCR gene amplification and sequencing, the authors found that there are about 10^6 different β chains in human blood and that each of these chains can bind, on average, to at least 25 different α chains. Therefore, a lower bound on the human TCR diversity was set to 2.5×10^7. A similar approach was used in Casrouge et al. (2000) to estimate

[1] In the context of this chapter, a *heteroduplex* is a double-stranded DNA (dsDNA) molecule composed of two imperfectly matching single strands. A *homoduplex* is a dsDNA molecule with perfectly matching strands.

the size of the TCR repertoire of naive mouse splenocytes. In these early studies, the use of mathematics was rather limited and modeling was not involved.

From a theoretical point of view, De Boer and Perelson (1993) developed a probability-based model in order to answer the question “How diverse should the immune system be?" Interestingly, the authors find that it is not the large number of foreign antigens that induces an enormous repertoire diversity, but rather the number of self antigens the immune system needs to avoid reactivity with. Later, Borghans et al. (1999) adopt a similar methodology to address the specificity of immunological memory. One of the conclusions of the authors is that memory lymphocytes should be more specific than the naive ones.

In the last decade, important advances in terms of experimental techniques have been achieved and with them, statistical and computational methods specifically applied to these techniques have started developing. Hitherto, the main data source seems to be the so-called high-throughput DNA sequencing (Mardis, 2008; Shendure and Ji, 2008). The latter is an optimized and more efficient version of the capillary-based sequencing used in the above-mentioned early studies (Arstila et al., 1999; Casrouge et al., 2000). T cells are sampled from the blood and are possibly sorted according to a surface receptor in order to isolate functional subsets, such as naive or memory cells (Wang et al., 2010). The DNA (or mRNA) coding for the complementarity determining region 3 (CDR3) of the TCR is then amplified and sequenced. The resulting data set contains the number of different DNA sequences and their frequency in the sample, i.e., all the information about the sample distribution. It is therefore straightforward to put the information in the desired form and, for example, estimate the clonotype or the clonal size distribution. An example of data representation can be found in Robins et al. (2009, Fig.6B), where the authors display a histogram of CDR3 sequences with certain length that use a certain gene segment. The availability of such detailed data sets made urgent the development of statistical estimators of the whole-repertoire diversity.

Estimating the diversity of TCRs is in fact a particular instance of a classical problem in statistics. The general setting is that of a population (TCRs) partitioned into a number of classes (clonotypes), each class corresponding to a distinct species. The goal is to estimate the number of different classes. Applications of that problem are found in numerous fields: in ecology for estimating the biodiversity of plants or animals, in computer sciences for database or social network searches, in linguistics for estimating the size of an author’s vocabulary (Efron and Thisted, 1976), and many others. Theoretically, there are several ways to tackle the problem, according to whether the population size is finite or infinite (see Bunge and Fitzpatrick (1993) for a review). In the context of TCR diversity estimation, the most common approach is to consider the “species sampling" process. The latter can be viewed as a stochastic process where new species “arrive" in a sample according to a Poisson process with a rate, itself varying according to a parametric (mixing) distribution. This approach (or other approximations of it) has been recently adopted by numerous groups (Rempala et al., 2010; Robins et al., 2009; Sepúlveda et al., 2010). Although the full TCR repertoire assessment remains an open question, using the above-mentioned statistical tools has led to refined diversity estimations. The latest of these (Robins et al., 2009) evokes a TCRβ receptor diversity at least 4-fold higher than previous estimates and 10-fold higher in the subset of antigen-experienced T cells.

Data output from other experimental techniques aiming the diversity assessment appear in other forms. For example, Immunoscope/Spectratype produces a histogram of CDR3 lengths, but gives no information about the diversity or the clonal size distribution. In AmpliCot, the distributional information is inherent to the data, but not directly available. Expanded clones are therefore more difficult to detect. This is why an additional modeling step is necessary to access all the information contained in the experimental data before extrapolating the whole repertoire diversity from a data sample.

4.1.2 Our Contributions

As mentioned earlier, the work presented here aims at the correct diversity estimation of a given repertoire sample. In the literature overviewed in the previous section, most of the theoretical and computational efforts focused on addressing the second step of the “general approach", namely inferring the whole-repertoire diversity from the measured sample diversity. The measurement technique was mainly DNA sequencing. Alternatives to high-throughput sequencing, such as AmpliCot, have not been widely used, partly due to the lack of methods for data analysis or to other experimental issues (Schütze et al., 2010). The main contribution of this chapter is the improvement of the interpretation of AmpliCot's experimental data. To our knowledge, this is the first time that a mathematical model is used to describe this particular experimental technique. As a result to our modeling, we show that the underlying model assumed in the original paper by (Baum and McCune, 2006) is insufficient to explain AmpliCot experimental data. A more detailed model, assuming heteroduplex- and transient duplexes formation, leads to significantly better data fits and accounts for the fluorescence loss observed in (Schütze et al., 2010). As a consequence, we show that the Cot-based method for interpreting the results of AmpliCot suggested in (Baum and McCune, 2006) should be applied with caution. Finally, we suggest alternative methods for sample diversity extrapolation.

4.1.3 Chapter Outline

This chapter is organized as follows. In the next section, among other methodological details, are presented the AmpliCot assay (Section 4.2.1), the DNA annealing kinetics models (Section 4.2.3) and the diversity extrapolation methods (Section 4.2.10). The results (Section 4.3) are then followed by a discussion (Section 4.4).

4.2 Materials and Methods

4.2.1 AmpliCot Assay

Samples containing PCR-amplified TCR genes of lymphocytes (or artificially synthesized oligonucleotides) were mixed with SYBR green fluorescent dye, which binds to double-stranded DNA. Aliquots of this mixture were placed in the upper and lower rows of a 96-well plate as the annealing sample and the reference (Figure 4.1A). A replicate (if any) was placed one row below the original sample. The pre-anneal step consists of measuring the baseline fluorescence of the samples and of the reference at annealing temperature (Figure 4.1B). Then, the temperature is increased for 2 minutes and the samples melt, whereas the reference stays at annealing temperature (melting step). Note that there is approximately a one-degree difference in the temperatures at which two replicates are melted, caused by the temperature gradient feature of PCR machines. The fluorescence intensity of the samples is strongly decreased during the melting step, as double-stranded DNA de-hybridizes. During the annealing step, the temperature of the samples was set back to annealing temperature and the fluorescence intensity as a function of time was measured every 5-20 seconds (Figure 4.1B). At equal concentrations, the re-annealing rate is dependent on the diversity of the sample.

4.2.2 Experimental data

Templates of known diversity were created from a library of individually synthesized oligonucleotides and are composed of 64 base pairs, out of which a maximum of 16 can differ. AmpliCot was performed on samples where all sequences are present in equimolar ratio. We had at our disposal three data sets: the original oligonucleotide data set of Baum & McCune (Baum and McCune, 2006, Fig.2a) and two other data sets obtained at the University Medical Center Utrecht (UMCU). To slow down the annealing kinetics, low diversity samples of UMCU data set 1 were diluted. The inverse of the dilution factor was used as a sample's concentration. For the other two data sets, the dilution (and therefore the concentration) was the same for all diversities. Table 4.1 summarizes the three data sets.

Data set	Diversities	Nb. of replicates	Dilution factor
Baum & McCune	$n = 1, 2, 5, 10, 30, 48, 96$	1	Same for all n
UMCU1	$n = 1, 4, 8, 16, 32, 48$	2 ($n = 1, 4, 8, 16$) 1 ($n = 32, 48$)	1:4 ($n = 1, 4$) 1:2 ($n = 8, 16, 32, 48$)
UMCU2	$n = 10, 20, 30, 40$	2	Same for all n

Table 4.1: The three data sets of known diversity templates at our disposal.

4.2.3 Modeling the DNA Annealing Kinetics

Second Order Kinetics (SOK)

Second order kinetics is the minimal model that describes the annealing phase of AmpliCot. It expresses the fact that two complementary single strands of DNA anneal and form a homoduplex. Consider a DNA sample of diversity n. Let S_i be the concentration of single-stranded DNA (ssDNA) molecules of type i, and D_{ii} be the concentration of homoduplexes of type i, $i = 1, \dots, n$. We call a the rate of association of two single-stranded molecules. Assuming a well mixed solution and a large number of molecules, the following differential equations describe the second order kinetics model (Figure 4.1C):

$$\begin{aligned} \frac{dS_i}{dt} &= -2aS_i^2 \\ \frac{dD_{ii}}{dt} &= aS_i^2. \end{aligned} \tag{4.1}$$

Let $t_0 = 0$ be the beginning of the annealing phase of AmpliCot and let T be the total concentration of ssDNA molecules inside a sample. Let f_i be the proportion of ssDNA of species i at the beginning of the annealing phase. Assuming that there is no loss of matter, if all dsDNA is denatured after the melting phase of AmpliCot, there would be f_iT single-stranded molecules of type i at the beginning of the annealing phase and no dsDNA. However, as the fluorescence level at melting temperature is not exactly 0, some molecules are possibly in double-stranded form. We call α the proportion of melted molecules at t_0 ($\alpha \in [0, 1]$). The initial conditions are thus $S_i(0) = \alpha f_i T$ and $2D_{ii}(0) = (1 - \alpha) f_i T$, $i = 1, \dots, n$. Let $F(t)$ be the concentration of fluorescent molecules at time t. In the case of second order kinetics,

$$F(t) = 2\sum_{i=1}^{n} D_{ii}(t) = T - \sum_{i=1}^{n} S_i(t).$$

Complete Model (CM)

This model describes in further detail the biochemical reaction of AmpliCot. We consider the fact that the hybridization involves two steps and it might result in heteroduplex formation because of erroneous associations (Figure 4.1C). Encounters of two single DNA strands occur at rate a. Two strands form either a partially hybridized homoduplex C_{ii} if they are both perfectly complementary, or a partially hybridized heteroduplex C_{ij} otherwise. Partially hybridized homoduplexes (resp. heteroduplexes) can dissociate at rate d_1 (resp. d_2), or hybridize completely at rate z_1 (resp. z_2) to form the final product D_{ii} (resp. D_{ij}). The differential equations describing the change in time of the above-mentioned concentrations are:

$$\begin{aligned}
\frac{dS_i}{dt} &= -2aS_i^2 - aS_i \sum_{j=i+1}^{n} S_j + 2d_1C_{ii} + d_2 \sum_{j=i+1}^{n} C_{ij} \\
\frac{dC_{ii}}{dt} &= aS_i^2 - (d_1 + z_1)C_{ii} \\
\frac{dC_{ij}}{dt} &= aS_iS_j - (d_2 + z_2)C_{ij} \\
\frac{dD_{ii}}{dt} &= z_1C_{ii} \\
\frac{dD_{ij}}{dt} &= z_2C_{ij},
\end{aligned} \tag{4.2}$$

where $i = 1, \ldots, n$ and $j = i+1, \ldots, n$. We assume that that the melting process is fast compared to the re-annealing, and that the melting temperature is so high that no re-hybridization is occurring during the melting phase. Under these hypotheses, the sample contains only ssDNA or unmelted dsDNA homoduplexes at the beginning of the annealing phase. The initial conditions for the above system are thus $S_i(0) = \alpha f_i T$, $D_{ii}(0) = (1-\alpha) f_i T$, $C_{ii}(0) = C_{ij}(0) = D_{ij}(0) = 0$, $i = 1, \ldots, n, j = i+1, \ldots, n$, $\alpha \in [0, 1]$. We assume that heteroduplexes have a decreased fluorescence intensity compared to homoduplexes and we model this by weighting their fluorescence by a factor $\varphi \in [0, 1]$. The concentration of fluorescent molecules is therefore defined as

$$F(t) = 2\left(\sum_{i=1}^{n} D_{ii}(t) + \varphi \sum_{i=1}^{n} \sum_{j=i+1}^{n} D_{ij}(t)\right).$$

4.2.4 Mean Field Models

An analytic solution for second order kinetics is easy to find, but this is not the case for the complete model. Solving numerically the above differential equations is computationally expensive because the size of the system grows with the diversity n. However, when the DNA particles of each species are in equimolar concentrations in the annealing buffer, the above differential equations can be simplified by using the fact that if $f_i = 1/n$, the dynamics of the whole system are proportional to the dynamics of the individual species. In other words, the dynamics of one specie are equal to the mean of all species. Using this property allows a reduction of the number of variables from $\mathcal{O}(n)$ (in the case of second order kinetics) or $\mathcal{O}(n^2)$ (in the case of the complete model) to $\mathcal{O}(1)$ (2 or 5 variables). This simplification is thus removing the dependence of the system size on the diversity. Hence, solving the differential equations numerically is no longer a hurdle because the computational cost of the numeric solution is constant. This is very important when dealing with samples of diversity in the orders of millions, such as, for example, DNA from human naive T cells. The derivation of the mean field equations are given in the following two sections.

Mean Field Second Order Kinetics

Let $S(t) = \sum_{i=1}^{n} S_i(t)$ and $D(t) = \sum_{i=1}^{n} D_{ii}(t)$. If all species are in equimolar concentrations in the mixture, we have $S_i(t) = S_j(t)$, $i \neq j$, $\forall t$. Therefore, $S(t) = nS_1(t)$, $D(t) = nD_{11}(t)$

and the differential equations Eq. (4.1) become

$$\begin{aligned}\frac{dS}{dt} &= -2a\frac{S^2}{n} \\ \frac{dD}{dt} &= a\frac{S^2}{n},\end{aligned} \tag{4.3}$$

where we have used the fact that $nS_1^2 = S^2/n$. The solution of this system with initial conditions $S(0) = \alpha T$ and $2D(0) = (1-\alpha)T$ is:

$$\begin{aligned}S(t) &= \frac{\alpha T}{1 + 2\frac{a}{n}\alpha T t} \\ 2D(t) &= T - S(t),\end{aligned}$$

and the fluorescent molecules

$$F(t) = 2D(t). \tag{4.4}$$

Mean Field Complete Model

Assuming equimolar concentrations of each species, we define the following quantities:

$$\begin{aligned}S(t) &= \sum_{i=1}^{n} S_i(t) = nS_1(t) \\ C(t) &= \sum_{i=1}^{n} C_{ii}(t) = nC_{11}(t) \\ H(t) &= \sum_{i=1}^{n}\sum_{j=i+1}^{n} C_{ij}(t) = \frac{n(n-1)}{2}C_{12}(t) \\ J(t) &= \sum_{i=1}^{n}\sum_{j=i+1}^{n} D_{ij}(t) = \frac{n(n-1)}{2}D_{12}(t) \\ D(t) &= \sum_{i=1}^{n} D_{ii}(t) = nD_{11}(t),\end{aligned} \tag{4.5}$$

where indices 1 and 2 have been chosen arbitrarily to design one species. $S(t)$ denotes ssDNA, $C(t)$, partially hybridized homoduplexes, $H(t)$, partially hybridized heteroduplexes, $J(t)$, final product heteroduplexes and $D(t)$, the final homoduplexes. The differential equations of Eq. (4.2) can be written in terms of the above variables as follows:

$$\begin{aligned}\frac{dS}{dt} &= -a(n+1)\frac{S^2}{n} + 2d_1C + 2d_2H \\ \frac{dC}{dt} &= a\frac{S^2}{n} - (d_1 + z_1)C \\ \frac{dH}{dt} &= a\left(\frac{n-1}{2}\right)\frac{S^2}{n} - (d_2 + z_2)H \\ \frac{dJ}{dt} &= z_2H \\ \frac{dD}{dt} &= z_1C,\end{aligned} \tag{4.6}$$

with initial conditions $S(0) = \alpha T, 2D(0) = (1-\alpha)T, C(0) = H(0) = J(0) = 0$, and fluorescent molecules

$$F(t) = 2(D(t) + \varphi J(t)). \tag{4.7}$$

This system of differential equations is solved numerically.

4.2.5 Modeling the AmpliCot Assay

Let $R(t)$ and $A_{\text{raw}}(t)$ be respectively the fluorescence intensity of the reference and of the sample at time t (in minutes). Let t_m be the start of the melting step and $t_0 > t_m$ be the start of the annealing step (Figure 4.1B). We adopt the following convention: $t_0 = 0$, which implies that $t_m < 0$. Let A_b be the baseline fluorescence of the sample, which we define as the fluorescence intensity at time 0 (Figure 4.1B). Similarly, let R_b be the baseline fluorescence of the reference. The following equations describe the pre-annealing and the annealing step of AmpliCot (we do not model the melting step, as it does not contain any pertinent information about diversity):

$$R(t) = R_b h(t) \tag{4.8}$$

$$A_{\text{raw}}(t) = \begin{cases} A_b h(t) & \text{for} \quad t < t_m \\ A_b F(t; \Theta, \alpha) h(t) & \text{for} \quad t \geq t_0, \end{cases} \tag{4.9}$$

where $F(t; \Theta, \alpha) \in [1 - \alpha, 1]$ is the concentration of fluorescent molecules determined by the biochemical reaction model (SOK (Eq. (4.4)), or CM (Eq. (4.7))), with reaction rates Θ, total concentration $T = 1$ and proportion α of melted molecules at time 0. The function $h(t)$ describes phenomenologically the slow fluorescence decay due to degradation of the SYBR green die (Baum and McCune, 2006). It has the following expression:

$$h(t) = \left(K_1 e^{-\delta_1 t} + K_2 e^{-\delta_2 t} + (1 - K_1 - K_2)\right), \tag{4.10}$$

where $K_1, K_2 \in [0, 1]$ and $\delta_1, \delta_2 > 0$.

4.2.6 Model Fitting to Experimental Data

In order to determine which of the above models of DNA annealing kinetics describes the experimental data best, we fitted each model to the three data sets. The mean field versions of the differential equations were solved numerically (except for the second order kinetics model, where an analytical solution is available) using a Runge-Kutta algorithm written in C. The parameters were fitted using a least squares procedure (implemented in Matlab) applied to the log-transformed raw annealing curves.

We assume that some parameters are common to all samples on the samples plate, whereas others are specific to each sample. We therefore splited the parameters in two groups: the common parameters Θ, and the sample-specific parameters Ω. The biochemical reaction rates and the fluorescence of heteroduplexes are common to all samples, whereas the fluorescence decay parameters (Eq. (4.10)) and the baseline fluorescence are specific to each sample. The proportion α of DNA molecules that are in single-stranded form at the beginning of the annealing phase is assumed to be common to all samples placed on the same row of the plate.

The replicates (if any) placed on another row were allowed to have a different α value, because the temperature at which samples are melted varies from one row to another[2] and therefore the amount of melted material observed after the same melting period can be slightly different. To simplify notations, we consider α as a common parameter and we make the distinction whenever unclear. To summarize, we have

$$\Theta = (a, d_1, d_2, z_1, z_2, \varphi, \alpha_j) \tag{4.11}$$

$$\Omega_i = (\delta_1^{(i)}, \delta_2^{(i)}, K_1^{(i)}, K_2^{(i)}, R_b^{(i)}, A_b^{(i)}), \tag{4.12}$$

where i denotes the i^{th} sample in a given experiment, $i = 1, \dots, N$, and j denotes the replicate index (if any), $j = 1$ or 2. The reference of each sample is used to fit δ_1, δ_2, K_1, K_2 and R_b. Note that if the pre-anneal measurements are long enough, they can be used as an alternative to fit the fluorescence decline parameters. The baseline fluorescence intensity of the sample (A_b) was estimated using the last 10 measurements of the pre-anneal step and corrected for the fluorescence decline during the melting step obtained from the reference. The annealing curves of all diversities were used to fit the common reaction rates Θ (Eq. (4.11)).

To slow down the annealing kinetics of samples with low diversities, some samples were diluted (cf. Table 4.1). The inverse of the dilution factor was used as an indicator of the concentration. Theorem 2 in Chapter 5 shows how dilution of a sample affects the reaction rates.

4.2.7 Confidence Intervals on Parameter Values

The 95% confidence intervals on parameter values were computed using 999 bootstrap replicates of each original data set. The bootstrap was done by sampling with replacement pairs of points $(t_i, A_{\text{raw}}(t_i))$ from pre-anneal and annealing curves. The bootstrap replicates where fitted in the same way as the original data set (see Section 4.2.6). To avoid that the optimization procedure gets trapped in the same local minimum when applied to bootstrap replicates, the initial point of the fitting procedure was randomized. The confidence intervals are computed using order statistics of the bootstrap distribution (Le Boudec, 2010).

4.2.8 Computing the Annealing Percentage (or Data Normalization)

In order to compute the percentage of annealed material, the raw sample was normalized. This was done by correcting for the baseline fluorescence discrepancies of the reference and the sample and by correcting for the time-dependent fluorescence decline. In addition, in order to obtain a percentage of annealed material in the range 0–1, the fluorescence of the sample at melting temperature (rescaled to the reference baseline, i.e., α in the notation of the model) was subtracted from the corrected annealing curve. Note that the latter transformation is only valid under the assumption that the fluorescence at melting temperature is due to measurement noise and is not an indication of annealed material. Let $A(t)$ be the resulting normalized annealing

[2] There is approximately one-degree difference in the in the temperatures at which two replicates are melted, caused by the temperature gradient feature of PCR machines.

curve expressed in "percentage annealed". We obtain $A(t)$ from the raw data for $t \geq t_0$ by applying the above transformations:

$$A(t) = \frac{\frac{A_{\text{raw}}(t)}{R(t)} - \frac{A_m}{R_b}}{\frac{A_b}{R_b} - \frac{A_m}{R_b}} = \frac{R_b}{A_b - A_m}\left(\frac{A_{\text{raw}}(t)}{R(t)} - \frac{A_m}{R_b}\right). \tag{4.13}$$

The above equation was used to normalize the data, where A_b and R_b were estimated by the mean of the last 10 measurements of the pre-anneal phase and A_m was estimated by the mean of the measurements during the melting step. If these measurements were not available, A_m was set to the first measurement of the annealing step. According to Baum and McCune (2006), the normalized data should be considered in concentration × time (Cot) units to account for the effect of the concentration differences in the annealing speed of samples. We have used the inverse of the dilution factor (whenever known) as an indicator of the concentration.

If the model is fitted on the raw data and then needs to be transformed in normalized form, the above equation can be further developed and simplified in order to be expressed in terms of the model parameters. Indeed, the fluorescence A_m of the sample after the melting reflects the concentration of unmelted material, which is expressed as $F(0) = 2(D(0) + \varphi J(0)) = 2D(0) = (1-\alpha)T$. A fluorescence intensity of A_b is associated to the total concentration T, therefore the fluorescence at melting temperature can be written as $A_m = f(F(0)) = (1-\alpha)A_b$. Thus, following the definition of $A_{\text{raw}}(t)$ (Eq. (4.9)) and of $R(t)$ (Eq. (4.8)), one can write:

$$\begin{aligned} A(t) &= \frac{R_b}{A_b - A_m}\left(\frac{A_b F(t;\Theta,\alpha)h(t)}{R_b h(t)} - \frac{A_m}{R_b}\right) \\ &= \frac{1}{\alpha A_b}\left(A_b F(t;\Theta,\alpha) - (1-\alpha)A_b\right)\right) \\ &= \frac{1}{\alpha}(F(t;\Theta,\alpha) - (1-\alpha)) \\ &\overset{\text{Th.1}}{=} F(t;\alpha\theta_1,\theta_2,1), \end{aligned} \tag{4.14}$$

where the last equality holds following Theorem 2 of the Appendix. $F(t)$ is defined in Eq. (4.4) or Eq. (4.7), depending on the model.

4.2.9 Cot$_p$ and t_p Values Estimation

In the original AmpliCot paper (Baum and McCune, 2006), to correct for concentration differences, time is multiplied by the nucleotide concentration of a sample (Cot = Concentration × time). From the normalized data, one can estimate the Cot value at which a proportion p of the sample has annealed ($p \in [0,1]$). This is what we call a Cot$_p$ value (equivalently, a "Cot X%" value). A t_p value is simply the time at which a proportion p of the sample has annealed ($p \in [0,1)$.

To estimate a t_p value from noisy data, we defined a precision interval around p, say $[p - \epsilon/2, p + \epsilon/2]$, where ϵ is the precision level. We considered the data points that fall inside the above interval. Let $\tilde{t}$ be the set of time coordinates of these points. As the annealing curve is monotonous and the measurement error small, it can be assumed that the search t_p value, say

t_p^*, is in $\tilde{t}$. If ϵ is small enough, the annealing curve is locally approximately linear. Therefore, taking simply $t_p^* \approx (\min(\tilde{t}) + \max(\tilde{t}))/2$ is approximately correct. Of course, this simple approximation is sensitive to aberrant values, but since the data is rather smooth and does not present big deviations from its mean value, this approximation should yield correct results. The same procedure was applied to Cot_p values.

4.2.10 Diversity Prediction Methods

A diversity prediction consists of analysis of samples of oligonucleotides with known diversity as a training set for calibration, then predicting the diversity of a sample with unknown diversity. We consider three different methods of prediction. The first method is the one suggested in Baum and McCune (2006), which we call the "Cot-based" prediction method. In short, the normalized annealing curves are considered at a single point (the Cot point at which X$\%$ of the material has annealed) and one assumes a linear relation between Cot X$\%$ and diversity n (Eq. (4.15)), as predicted by second order kinetics. The slope of the latter relation is estimated using a linear regression and the unknown diversity is predicted from the inverse relation evaluated at the Cot X$\%$ point of the unknown sample.

We propose two alternative methods. The "model-based" method intends to use not only the information of the annealing curves sampled at one point as in the previous procedure, but the information contained in the entire annealing curve. It also has the advantage of having the choice of using the raw or the normalized data. The "model-based" prediction uses the samples of known diversity to fit the parameters of the underlying annealing kinetics model (second order kinetics or the complete model). The biochemical reaction rates are then fixed to their best-fit values and a range of possible values of the diversity n is determined. For each value of this pre-determined set, the remaining parameters are estimated by fitting the model to the sample with unknown diversity by minimizing the least-squared errors of the log-transformed data and model. The log-likelihood of the resulting fit is calculated. The unknown diversity is estimated as the value of n that maximizes the log-likelihood.

The third method, called "t_p-based", is an improved version of the above Cot-based method. It uses t_p values, but instead of assuming a linear relation between t_p and n, it uses the best-fit of the model to infer the (possibly nonlinear) relation. The unknown diversity is read from the inverse relation. Table 4.2 summarizes the three prediction methods.

Note that the Cot-based method is assuming second order kinetics as the underlying model for the annealing kinetics. The other two methods can be applied assuming one or the other underlying models, i.e., second order kinetics or the complete model. However, for the rest of the manuscript, we will apply both the model- and the t_p-based methods assuming the complete model.

4.2.11 Confidence Intervals on Predictions

In order to test the accuracy of the above prediction methods, we used the low diversity samples to execute the calibration step of the procedures and we predicted one of the high diversity samples ($n = 48$ or 96, depending on the data set). A confidence interval on the prediction was computed by bootstrapping the diversities on which the model was calibrated

Method	Calibration	Prediction
Cot-based	**1.** Normalize the data.	**1.** Estimate the Cot_p of the unknown diversity sample.
	2. Choose a value of $p \in [0, 1]$.	**2.** Return the value n^* at which the linear regression is equal to the estimated Cot.
	3. Estimate Cot_p values for each known diversity.	
	4. Do a linear regression on the Cot_p values as function of n.	
Model-based	**1.** Fit the common parameters Θ (Eq. (4.11)) of the model (second order kinetics or the complete model) to the raw data of the known diversity samples.	**1.** Given a value of $n \in \tilde{N}$ and the best-fitting common parameters Θ (Eq. (4.11)), fit the remaining parameters Ω (Eq. (4.12)) to the raw data of the unknown diversity sample and compute the likelihood of the resulting fit.
	2. Define a large set $\tilde{N}$ of possible diversities susceptible to contain the unknown diversity.	**2.** Repeat step 1. for all $n \in \tilde{N}$.
		3. Return the diversity n^* with the highest likelihood.
t_p**-based**	**1.** Fit the common parameters Θ (Eq. (4.11)) of the model (second order kinetics or the complete model) to the raw or normalized known diversity samples.	**1.** Estimate the value $\tilde{t_p}$ of the unknown diversity sample by normalizing it.
	2. Normalize the data and the model (if not fitted on normalized data).	**2.** Return the value n^* at which the (nonlinear) inferred $t_p(n)$ relation is equal to $\tilde{t_p}$.
	3. Choose a value of $p \in [0, 1]$.	
	4. Based on the best-fit parameters, infer the relation $t_p(n)$ for large n.	

Table 4.2: Prediction methods. The Cot-based method is the one suggested in (Baum and McCune, 2006) and is based on the assumption that second order kinetics is the correct underlying model to describe AmpliCot experiments. The Model-based method uses any underlying model that is fitted to the full data set; the best-fit parameters are then used to extrapolate an unknown diversity using the entire annealing curve (not only sampled at one point as in the Cot-based method). The t_p-based method is a hybrid of the other two: it considers Cot values, but does not assume a linear relation between Cot and n.

(at least 500 bootstrap replicates). The diversity of the unknown sample was estimated for each set of bootstrapped calibration curves and statistics were performed on the resulting bootstrap prediction distribution.

4.3 Results

4.3.1 Two Models Describing AmpliCot

The aim of AmpliCot is to measure the diversity of a pool of DNA sequences and has been applied to measure TCR diversity (Baum and McCune, 2006). Samples containing PCR-amplified products (be it from TCR or BCR genes of lymphocytes or artificially synthesized oligonucleotides) are mixed with SYBR green fluorescent dye, which binds double-stranded DNA (dsDNA). Aliquots of this mixture are placed in the upper and lower rows of a 96-well plate as the annealing sample and the reference (Figure 4.1A). A replicate (if any) is placed one row below the original sample. The AmpliCot procedure is comprised of three steps: the pre-anneal, melting and annealing step (Figure 4.1B). The pre-anneal step is intended to measure the baseline fluorescence of samples and references at annealing temperature. Preliminary experiments are performed in order to determine the optimal annealing temperature. In many cases, it is set to three degrees less than the melting temperature T_m. During the melting step, the temperature is increased to T_m for two minutes, which leads to the dissociation of dsDNA. By breaking its hydrogen bonds, dsDNA falls apart into two single-stranded DNA molecules. As a consequence, the fluorescence intensity of the sample strongly decreases. During the annealing step, samples are put back at annealing temperature and their fluorescence slowly increases as ssDNA molecules re-associate. Fluorescence readings are performed every 5–20 seconds, depending on the experiment. At equal concentrations, the re-annealing rate is dependent on the diversity of the sample.

We consider two models describing the biochemical reaction of the annealing step of AmpliCot: *second order kinetics (SOK)* and the *complete model (CM)* (Figure 4.1C). We assume that samples contain a large amount of DNA and that the material is well-mixed, so both models are defined using ordinary differential equations. The main difference between the models is the level of detail incorporated in the description of the underlying biochemical reaction.

Second order kinetics is the simplest model that allows the description of AmpliCot. It relates the association of two perfectly complementary single DNA strands under the assumption that the encounter of two strands is the rate limiting step and that the subsequent polymerization is fast compared to the former process. Under these assumptions, the hybridization of DNA is indeed a second order reaction. The differential equations describing SOK can be found in Eq. (4.1). This model was proposed in the original AmpliCot paper (Baum and McCune, 2006).

The complete model takes into account the fact that hybridization involves two distinct processes: the association of short, homologous sites on two single strands, followed by a reversible polymerization (Wetmur and Davidson, 1968). The partially associated complexes can either fall apart or definitely associate. The model also accounts for the possibility of heteroduplex formation which happens in two steps, similarly to homoduplex formation. A detailed description of the complete model and the differential equations describing it (Eq. (4.2)) can be found in the Materials and Methods (Section 4.2.3).

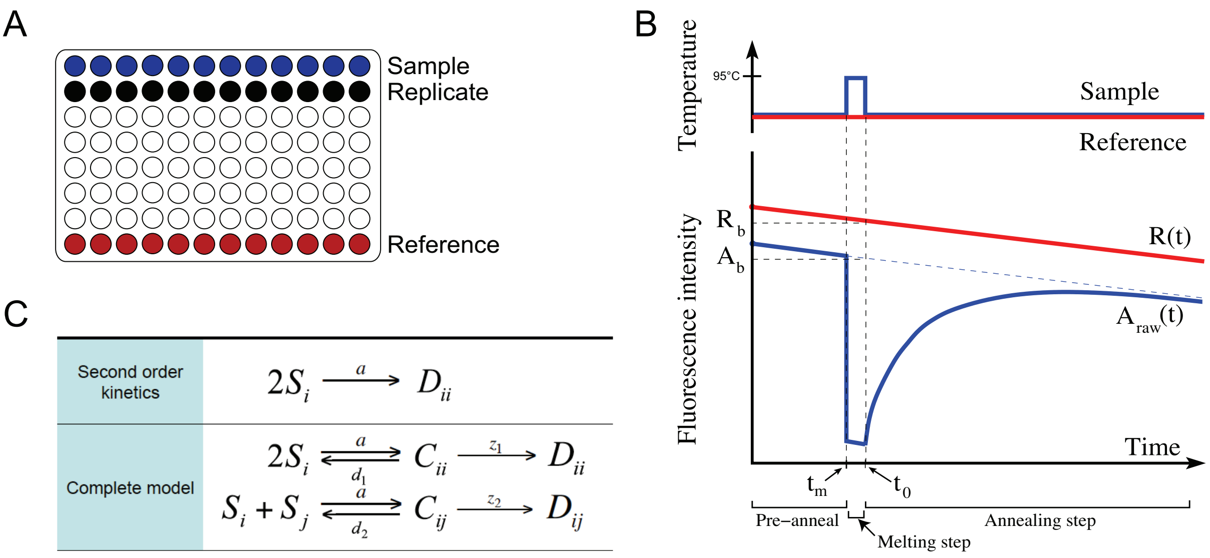

A
Sample
Replicate
Reference
B
Temperature
95°C
Sample
Reference
Fluorescence intensity
R_b
A_b
R(t)
$A_{raw}(t)$
Time
t_m
t_0
Pre−anneal
Annealing step
Melting step
C
Second order kinetics
$2S_i \xrightarrow{a} D_{ii}$
Complete model
$2S_i \underset{d_1}{\overset{a}{\rightleftharpoons}} C_{ii} \xrightarrow{z_1} D_{ii}$
$S_i + S_j \underset{d_2}{\overset{a}{\rightleftharpoons}} C_{ij} \xrightarrow{z_2} D_{ij}$

Figure 4.1: The AmpliCot assay and model. A: Samples containing PCR-amplified TCR genes or oligonucleotides are placed on both extremities of a 96-well plate as the samples and the reference. B: Data collection in a real-time PCR machine. The baseline fluorescence intensity of samples and reference is measured at annealing temperature (pre-anneal step). The sample is then melted at 95°C and its fluorescence drops (melting step). After 2 min. of melting, the temperature of the sample is quickly set back to annealing temperature to allow for re-annealing of DNA strands (annealing step). C: Two possible models of the biochemical reactions occurring during the annealing step of AmpliCot. Second order kinetics (top line) is the minimal model in which only homoduplexes are formed. The complete model (bottom line) considers the reaction in more details. The association occurs in two steps, a first encounter, followed by a zippering reaction. It also includes the possibility of heteroduplex formation.

4.3.2 The SOK model gives a good description of Baum&McCune's data

In order to evaluate which model best reflects the experimental reality, we have fitted both models to the raw data of Baum and McCune (2006). For details on the fitting procedure, we refer the reader to Section 4.2.6 of the Materials and Methods.

Figure 4.2 illustrates the experimental data of the pre-anneal ($t < 0$) and the anneal ($t \geq 0$) steps of AmpliCot, together with the best-fit of both models (green: SOK, red: CM). Each panel corresponds to a sample of one particular diversity. Visually, both models fit the data reasonably well and the maximum likelihoods of both fits are rather similar (Table 4.3). Although the complete model gives a significantly better fit to the data compared to second order kinetics (according to the F-test for nested models, p-value $< 10^{-3}$), the latter, which has 6 parameters less, is producing a fairly good description of the data. In addition, the best-fit parameters of the complete model (Table 4.3) suggest that the rate limiting step of the reaction is the first association and not the subsequent polymerization (z_1 is large compared to a). Since the best-fit parameters suggest that homoduplex formation does not necessitate transient complexes ($d_1 \approx 0$ and z_1 large), we tested whether the complete model can be simplified in that direction by fixing d_1 to 0 and z_1 to a very large value (1000). It resulted in a model version where only

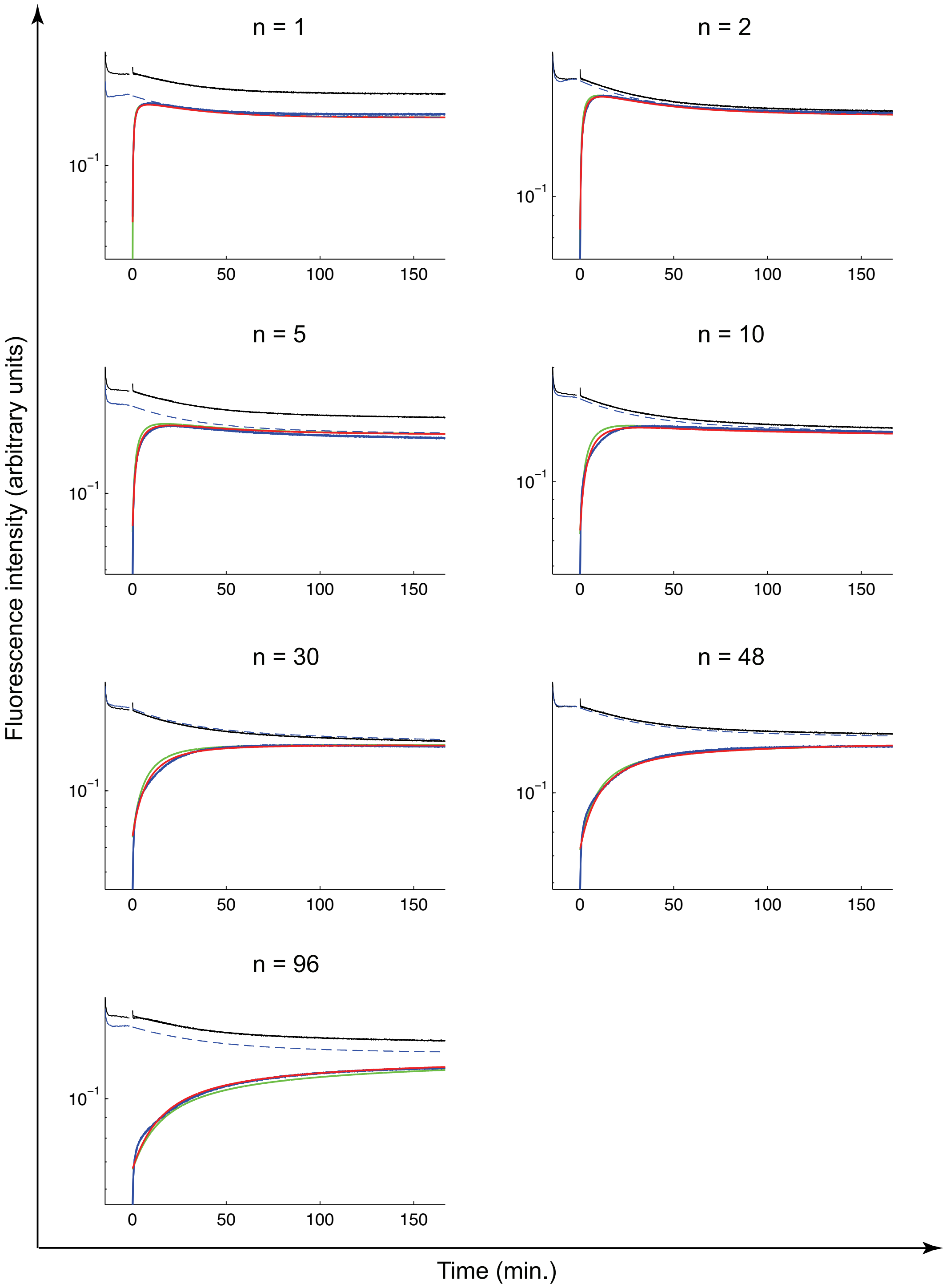


Figure 4.2: Best-fit of Baum&McCune's raw data for known diversity templates (each panel). Solid blue: data sample ($t < 0$: pre-anneal phase, $t \geq 0$: annealing kinetics). Dashed blue lines: the inferred kinetics of the "unmelted" sample, as predicted by the model. Black: reference. Solid green: best-fit of second order kinetics. Solid red: best-fit of the complete model. The fit of Eq. (4.8) to the reference is not visible, because it perfectly overlays the data. For color references, please consult the online electronic version of the thesis.

Model	Param.	Data set					
		Baum		UMCU1		UMCU2	
		Value	95% CI	Value	95% CI	Value	95% CI
SOK	ML	35 663		11 241		6 888	
	a	2.45	[2.42, 2.47]	13.36	[13.26, 13.41]	7.59	[7.5695, 7.6097]
	α_1	0.5522	[0.5513, 0.5530]	0.7334	[0.7327, 0.7340]	0.8048	[0.8043, 0.8051]
	α_2	-	-	0.8110	[0.8103, 0.8115]	0.7948	[0.7941, 0.7951]
CM	ML	39 444		14 532		13 987	
	a	2.06	[1.81, 7.38]	509.66	[454.1, 920.84]	38.77	[34.54, 47.84]
	d_1	0.0058	[0.00, 90.09]	73.2	[66.2, 134.4]	952.88	[640.04, 1500.00]
	d_2	7.14	[5.55, 656.91]	925.5	[821.2, 1767.7]	59.68	[53.32, 73.63]
	z_1	27.5	[21.1, 132.9]	2.67	[2.64, 2.73]	161.30	[122.64, 216.22]
	z_2	0.099	[0.079, 8.079]	1.69	[1.63, 1.79]	1.60	[1.58, 1.62]
	φ	0.9921	[0.9773, 0.9999]	0.8195	[0.8142, 0.8243]	0.8762	[0.8752, 0.8770]
	α_1	0.5514	[0.5496, 0.5639]	0.7598	[0.7592, 0.7603]	0.8698	[0.8696, 0.8702]
	α_2	-	-	0.7679	[0.7672, 0.7686]	0.8629	[0.8626, 0.8633]
	$\frac{d_1}{(d_1+z_1)}$	0.0002	[≈ 0, 0.7086]	0.9648	[0.9608, 0.9803]	0.8552	[0.8383, 0.8822]
	$\frac{d_2}{(d_2+z_2)}$	0.9863	[0.9851, 0.9961]	0.9982	[0.9979, 0.9990]	0.9739	[0.9707, 0.9788]

Table 4.3: Best-fit reaction rates and 95% confidence intervals (CI). Each model was fitted to the data by minimizing the sum of squared errors (on log scale). ML is the maximum likelihood of the best-fit. The CI (in parenthesis next to the parameter value) were computed using 999 bootstrap replicates (Carpenter and Bithell, 2000). Because there is a tradeoff between the values of d_1 and z_1 (resp. d_2 and z_2), the fractions $\frac{d_1}{(d_1+z_1)}$ and $\frac{d_2}{(d_2+z_2)}$ are shown for easier comparison of the three data sets.

heteroduplexes are formed in a two-step process (via transient complexes). The quality of the fit of the simplified model with respect to the complete model was not altered (p-value ≈ 1). Thus, for Baum's data set, the complete model is not minimal. Furthermore, only ~ 1% of the formed complexes end up in heteroduplex form and the fluorescence of the latter is only slightly lower than that of homoduplexes ($\varphi = 0.99$). We therefore conclude that second order kinetics is good enough to describe this data set.

Note furthermore that the confidence intervals of the bio-chemical rates of the complete model are fairly large (Table 4.3), indicating that several combinations of parameters produce good fits for this model. As the largest variability is noted for parameters d_1, d_2, z_2 and z_2, we believe that the values of these parameters are correlated (any possibly not identifiable separately). We therefore provided the values and confidence intervals for $\frac{d_1}{(d_1+z_1)}$ and $\frac{d_2}{(d_2+z_2)}$, which are more robust. The latter quantities confirm that d_2 is always large compared to z_2 (CI for $\frac{d_2}{(d_2+z_2)}$ does not include 0.5). However, d_1 is not always smaller than z_1, as the confidence interval for $\frac{d_1}{(d_1+z_1)}$ includes 0.5. This indicates that the complete model is still an option to be preferred over the simplified version of the previous paragraph.

So, if second order kinetics are a good approximation, the Cot-based prediction procedure (which assumes the validity of SOK) should yield reasonable results, at least when applied to the extrapolation of an unknown diversity close to the diversities of the calibration set. In order to make predictions using the Cot-based method suggested in Baum and McCune (2006), the raw data were normalized by applying the transformation of Eq. (4.13) and the value of $\text{Cot}_{0.5}$ was estimated from the data and plotted as a function of the diversity n. The resulting plot (Figure 4.3) is a reproduction of Figure 2c of (Baum and McCune, 2006). The low diversity samples ($n = 1, 2, 5, 10, 30, 48$) are used to make a linear regression. The Cot 50% value of the sample with "unknown" diversity ($n = 96$) is then read on the y-axis of the plot and the value of n at the intercept with the regression line is returned as the searched diversity. The confidence interval on the prediction is computed here from the confidence limits of the regression. When the method is calibrated on the samples with diversities 1-48 (Figure 4.3A), the result of the Cot-based prediction is $n = 77$, which is a $\sim 20\%$ underestimation of the true diversity. The confidence interval ($[69, 87]$) is rather large and it is not far from containing the true value.

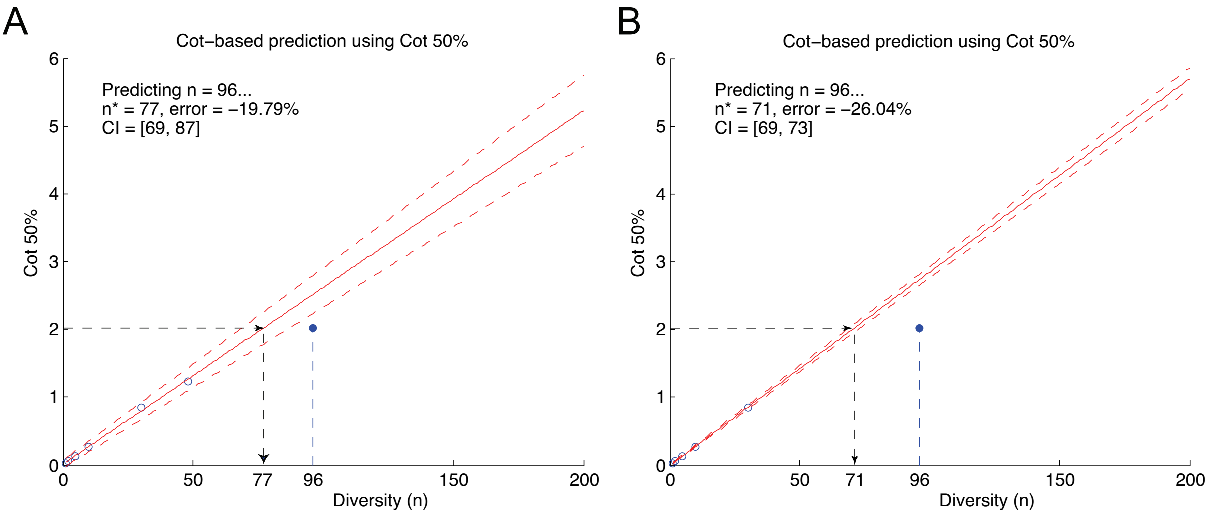


Figure 4.3: Cot-based prediction using Cot 50% of Baum&McCune's data. The raw data has been normalized using Eq. (4.13) and the value of Cot 50% is plotted as function of diversity. A linear regression (red solid line) is performed using the calibration set (○) and the intercept of Cot 50% of the "unknown" sample (●) with the regression line is used to make a diversity extrapolation (prediction). Dashed red lines: confidence interval of the regression. A: Calibration on $n = 1, 2, 5, 10, 30, 48$ gives a prediction of $n = 77$ (instead of 96), -20% error. B: Calibration on $n = 1, 2, 5, 10, 30$ gives a prediction of $n = 71$, -26% error.

Nevertheless, the practical application of AmpliCot is to measure the diversity of T- or B-cell receptors, ranging in millions, by calibrating on a library of known diversities, orders of magnitude smaller. We are therefore interested in assessing the performance of the prediction methods in conditions as close as possible to AmpliCot's practical application. Hence, as a toy example, we apply the Cot-based method to predict the highest diversity ($n = 96$) by calibrating on samples with diversities 1-30, thus omitting $n = 48$ (Figure 4.3B). The error of the resulting prediction has increased to 26% below the true value, while the confidence interval has shrunk ($[69, 73]$). Thus, the Cot-based method leads to underestimating the unknown diversity and the error is increasing with the gap between the most diverse calibration sample and the unknown diversity. This questions the linearity of Cot values as function of n and consequently, the use of the Cot-based prediction method in practical situations.

4.3.3 The bias introduced by the use of SOK instead of CM has an important effect on predicting large diversities

We saw in the previous section that even though second order kinetics yielded a reasonable description of the data, the Cot-based prediction method, which is based on SOK, failed to provide correct diversity extrapolations. The systematic underestimation of the true diversity questioned the linearity of Cot vs n values. Intuitively, the underestimation suggests that the relationship is concave. If this is indeed the case, even a slight deviation from linearity on the scale of 1–100 diversities could introduce an important bias when extrapolating large diversities.

Having in mind that the CM gave a significantly better fit of the data, we investigated the shape of the $t_p(n)$ function under this model [3]. To this end, we applied the transformation of Eq. (4.14) to the best-fit of the complete model, we estimated $t_{0.5}$ and plotted it against n (Figure 4.4A, •). For comparison, we also plotted on the same figure $t_{0.5}$ of the data (◆) and of the SOK model (▪). As expected from theory, SOK exhibited a linear $t_p(n)$ relation, whereas linearity was lost under the CM. Note that the deviation from linearity is small on the scale of diversities included in Baum's data set ($n \in [1, 100]$).

The initially small bias is however strongly amplified when considering large diversities. Figure 4.4B reveals that using one model, instead of the other, results in completely different functions. In other words, the same $t_{0.5}$ value has a completely different diversity-response of the inverse $t_{0.5}(n)$ curve. Thus, this confirms the intuition that if the annealing kinetics are better approximated by the complete model, using the linear relation of second order kinetics would largely underestimate the searched diversity. The next two sections describe the consequences of the validity of the complete model.

4.3.4 Under the complete model, Cot scaling does not correct for concentration differences

We address here the issue of concentration differences between samples. To understand the problem, consider two samples of same diversity, say $n = 1$ for simplicity. Imagine that one of them is twice the concentration of the other. Intuitively, in the more concentrated sample, molecules need to cover a smaller distance before meeting another molecule, so we expect that a higher concentration would result in a faster annealing.

Mathematically, the expression defining the annealing curve (Eq. (4.4)) under second order kinetics contains the product Tt, i.e., the total concentration of DNA × time. This implies that for a given diversity, the same experiment performed with two different concentrations will result in different annealing curves. In other words, two identical annealing curves plotted in

[3] Note that we consider t_p and not Cot values here, because Cot scaling can not be applied to the CM, as shown in Section 4.3.4.

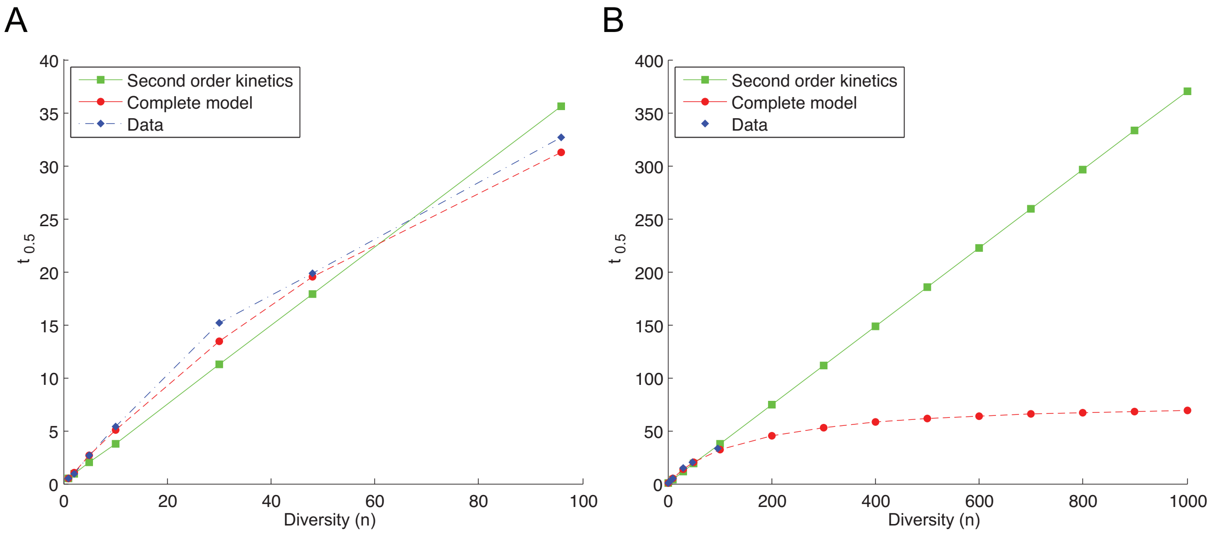


Figure 4.4: Using the Cot-based prediction method that assumes SOK when reality is better approximated by the complete model leads to underestimating large diversities. A: The behavior of Cot 50% as function of n is computed using both models: second order kinetics (■ solid) and the complete model (● dashed). The parameters of the best-fit of Baum's raw data have been used. Cot 50% values estimated from the normalized data are also plotted (◆ dash-dotted). For small diversities, the deviation from linearity is small. B: For larger diversities, the small bias between both models is amplified, leading to possibly wrong diversity predictions. Connecting lines are shown to help the visualization of the trend.

time scale do not necessarily imply the same sample diversity. This is a serious problem when one wants to predict the diversity of a sample out of its annealing kinetics. The Cot scaling used in (Baum and McCune, 2006) is a solution to this problem in the case of SOK, but it does not apply under the complete model (see Section 5.2 of Chapter 5 for a mathematical justification).

To illustrate the effect of a concentration change and Cot re-scaling, we have performed *in silico* experiments using both models. Figure 4.5A depicts the annealing kinetics of two virtual samples, one having total nucleotide concentration of $T_1 = 1$ (solid) and the other, concentration $T_2 = 2$ (dashed). The sample with highest concentration anneals faster, because molecule encounters occur more frequently. The annealing curve of sample 2 (in time scale) is therefore increasing faster than the one of sample 1 (Figure 4.5A) although both samples have the same diversity ($n = 40$). In order to recover the same annealing curve for both samples, the annealing kinetics of sample 2 should be slowed down by multiplying time with the concentration. The result of Cot re-scaling is shown in Figure 4.5B, where both curves overlay perfectly.

In the case of the complete model, a Cot re-scaling can not compensate for concentration differences, because the functions defining the ODE system of this model are not homogeneous[4]. A numerical example is shown in Figure 4.5C and D where the *in silico* experiment was performed under the complete model. The Cot-scaled curves of Figure 4.5D do not overlay

[4] A homogeneous function is a function with multiplicative scaling behavior: if the argument is multiplied by a factor, then the result is multiplied by some power of this factor. See Section 5.2 of Chapter 5 for a mathematical illustration.

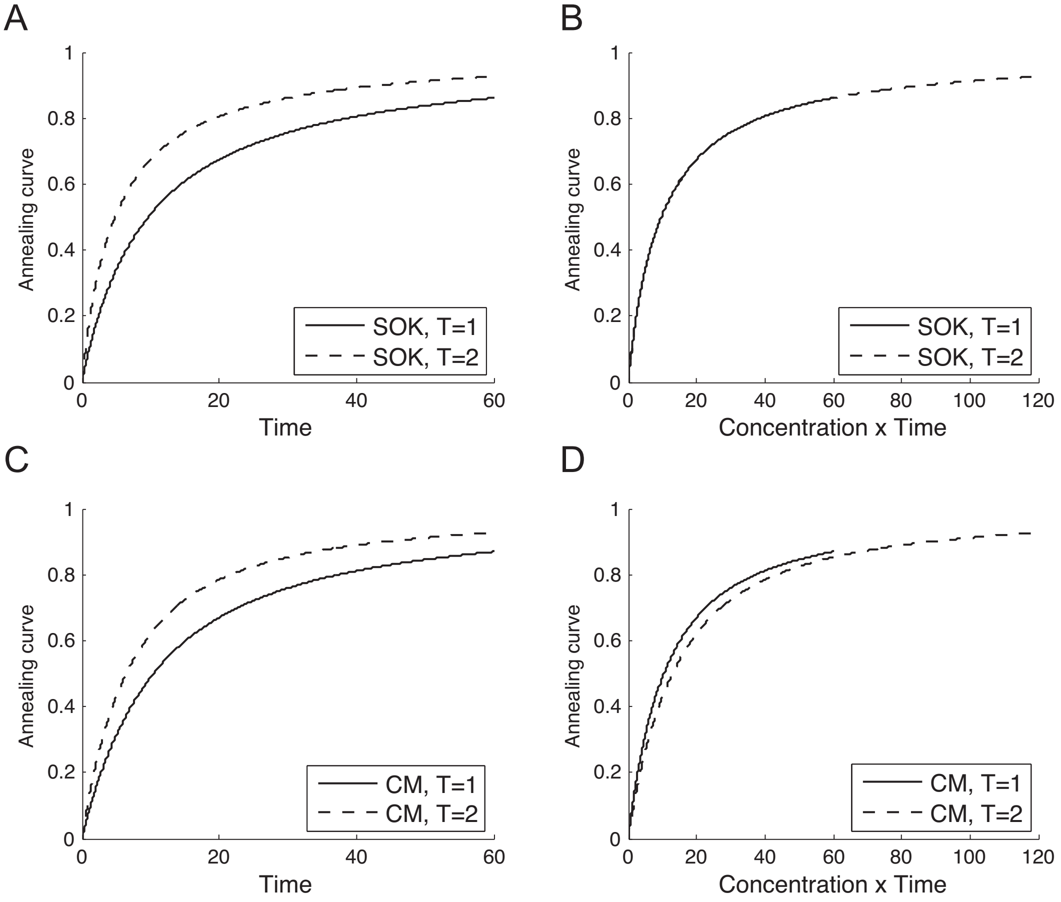


Figure 4.5: Cot scaling is valid under second order kinetics, but can not correct for differences in concentrations under the complete model. Annealing curves for two *in silico* experiments having different concentrations ($T_1 = 1$ and $T_2 = 2$). Top two panels: second order kinetics plotted on a time scale (A) and on a Cot scale (B). The curves in B overlay perfectly, showing that the Cot scale corrects for the concentration differences. Bottom two panels: heteroduplex model on a time scale (C) and on a Cot scale (D). In this case, Cot scaling does not lead to an overlapping of the two curves. Parameters: best-fit parameters of Baum's data (Table 4.3), $n = 40$.

perfectly as was the case under SOK. This is due to the fact that time does not appear solely in a product with the concentration in the solution of the complete model.

The correct way to handle concentration differences that is valid under both models, is by rescaling the association rates, as shown in Theorem 2 of Chapter 5.

4.3.5 Under the complete model, the function $t_p(n)$ is not linear

Here, we formalize mathematically the intuition of Section 4.3.3. The result presented here applies also to Cot values under second order kinetics, but because Cot scaling is not valid under the complete model, we derive it for t_p values. Consider t_p, $p \in [0, 1]$, the time point at which a proportion p of the sample has annealed. From the definition of the annealing curve (Eq. (4.14)), the latter is expressed as $A(t_p) = p$. In the case of second order kinetics, t_p is a linear function of diversity, as stated in Baum and McCune (2006). Indeed,

$$t_p(n) = \frac{1}{2a\alpha}\left(\frac{p}{1-p}\right)n, \quad n \in \mathbb{N}. \tag{4.15}$$

This above expression was plotted in Figure 4.4 ($p = 0.5$).

The linear relationship is lost under the complete model, as we saw in Figure 4.4 using the numerical solution of the ODEs (Eq. (4.6)). An analytical expression of $t_p(n)$ under the CM is available in the particular case of a quasi-steady state (QSS) assumption. If the values of the reaction rates are such that the transient complexes C and H can be assumed in steady state, the ODE system is then easily solved (see Section 5.3.4 of Chapter 5 for a solution). From the resulting solution, one can derive the following expression:

$$t_p(n) = \frac{pn}{K_1 + K_2(n-1)}, \quad n \in \mathbb{N},$$

where

$$\begin{aligned} K_1 &= 2\alpha a\left(\frac{z_1}{d_1 + z_1}\right)(1-p), \quad K_1 > 0 \\ K_2 &= \alpha a\left(\frac{z_2}{d_2 + z_2}\right)(\varphi - p), \quad K_2 \in \mathbb{R}. \end{aligned}$$

Note that $t_p(n)$ is a rational (non-linear) function of n, except when $p = \varphi$. In that particular case, $K_2 = 0$ and $t_p(n) = pn/K_1$ is again linear, but this is not generally true.

4.3.6 Fitting other data sets suggests that SOK is not valid

In order to verify the validity of both models on other data sets, additional experiments were performed. We present two data sets, one containing diversities $n = 1, 4, 8, 16, 32, 48$ and the second, $n = 10, 20, 30, 40$. The fits of both models to these data sets are depicted in Figure 4.6 (data set 1) and Figure 4.7 (data set 2). The improvement brought by the complete model was significant (p-value < 10^{-3}) and this is especially clear when looking at the best-fit of the higher diversity samples. We could clearly distinguish the time course of both models for $n = 32$ and $n = 48$ of data set 1, where SOK was unable to reproduce the correct curvature and the apparent “limit" value of the data. Moreover, SOK failed to give the correct asymptote value in the fit of UMCU data set 2.

The best-fit parameters of the CM (Table 4.3) are very different from those of Baum’s data. For UMCU data set 1, the association and dissociation rates (a, d_1, d_2) of transient complexes are large compared to the polymerization rates z_1 and z_2. This suggests that a quasi-steady state assumption for variables C and H can be applied to this data set. In this case, an analytical solution of the CM is available (see Section 5.3 of Chapter 5 for its derivation). For UMCU data set 2, the QSS assumption can be valid too, but it seems less obvious in that case. In order to verify if the assumption is applicable, one should compute and formally compare the two time scales of the model, as done in Borghans et al. (1996).

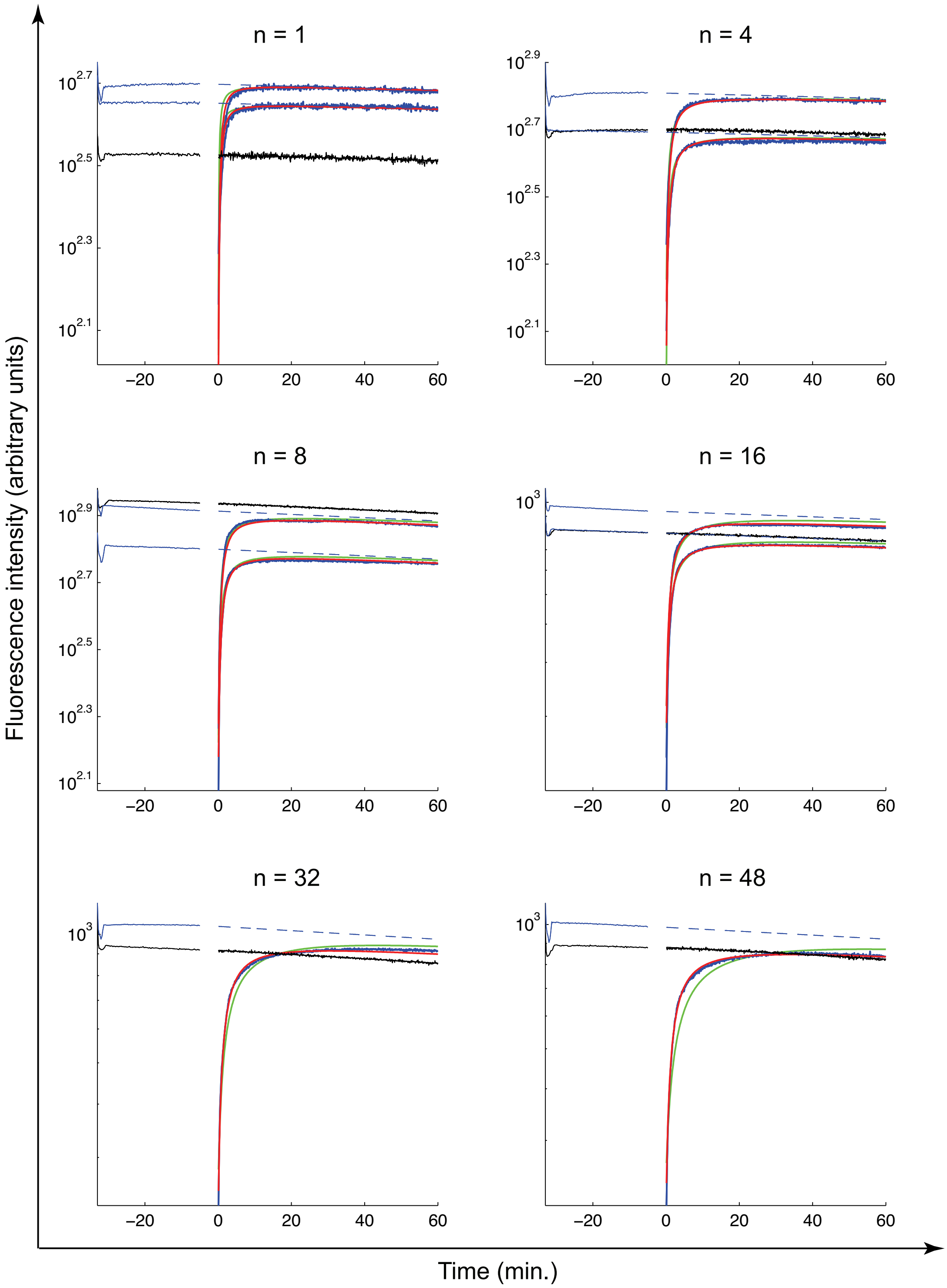


Figure 4.6: Best-fit of UMCU data set 1 for known diversity templates (each panel). Solid blue: data sample ($t < 0$: pre-anneal phase, $t \geq 0$: annealing phase). Dashed blue: the inferred kinetics of the "unmelted" sample, as predicted by the model. Black: reference. Solid green: best-fit of SOK. Solid red: best-fit of the CM. For color references, please consult the online electronic version of the thesis.

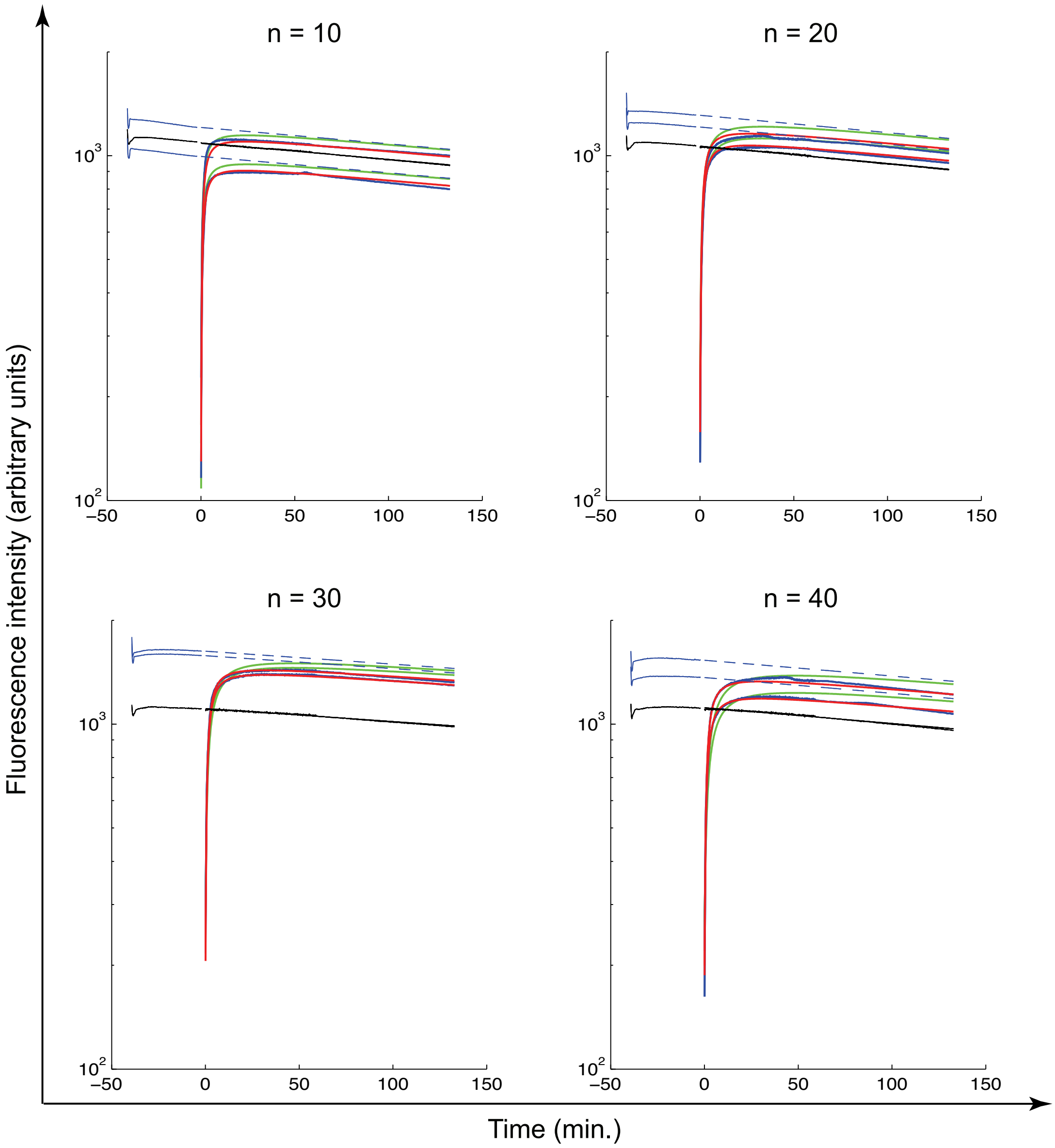


Figure 4.7: Best-fit of UMCU data set 2 for known diversity templates (each panel). Solid blue: data sample ($t < 0$: pre-anneal phase, $t \geq 0$: annealing phase). Dashed blue: the inferred kinetics of the "unmelted" sample, as predicted by the model. Black: reference. Solid green: best-fit of SOK. Solid red: best-fit of the CM. For color references, please consult the online electronic version of the thesis.

Similarly to Baum's data, we remark that the confidence intervals of the composite parameter $\frac{d_2}{(d_2+z_2)}$ are very tight around 0.99 for UMCU1 and around 0.97 for UMCU2. Hence, d_2 is always large with respect to z_2, indicating that a large proportion of complexes fall apart before definitely associating. Contrary to Baum's data, the same is true for d_1 and z_1 of UMCU data sets (d_1 is always large with respect to z_1).

Note that we have fitted two values for the proportion of single-stranded material at the start of the annealing (α_1 and α_2), each corresponding to samples of the same row. α_1 corresponds to the samples placed on the extremity of the plate, i.e., those that were melted at the highest temperature. We therefore expect that in these samples, the material melts somewhat faster, which would lead to a higher α value ($\alpha_1 > \alpha_2$). This is indeed the case for UMCU2 data set, but not for UMCU1. In fact, even though both best-fit values are very close to each other ($\alpha_1 = 0.7598$, $\alpha_2 = 0.7679$), their confidence intervals do not overlap which indicates that both values are significantly different. This counter-intuitive result, together with the fact that the best-fit values of α_1 and α_2 are very close to each other, suggests that the model can be further simplified by imposing one single α parameter for all replicates.

Finally, the best-fit parameters indicate that the fluorescence of heteroduplexes was estimated to be only 80-90% of that of homoduplexes ($\varphi = 0.82$ for data set 1, $\varphi = 0.89$ for data set 2), which contrasts with Baum's data where the few heteroduplexes could have almost the same fluorescence as homoduplexes.

4.3.7 Diversity-dependent fluorescence loss

We discuss here a phenomena that is typically observed in AmpliCot assays. Once the annealing has been completed, it has been often remarked that samples do not reach their pre-anneal values (even after correction for the slow fluorescence decay). This leads to a *gap* as shown in the example of Figure 4.8A. We define the gap as the difference between the inferred unmelted sample and the "limit" values of the annealing samples, predicted by the model. By “limit" values, we mean that the model is executed for a long time-period (for example 10 hours) with the best-fit parameters of the corresponding data set. Thus, the model is used to simulate a very long experiment in which all the material re-anneals ($t \to \infty$). A diversity-dependent gap is then observed in the three data sets (Figure 4.8B and C).

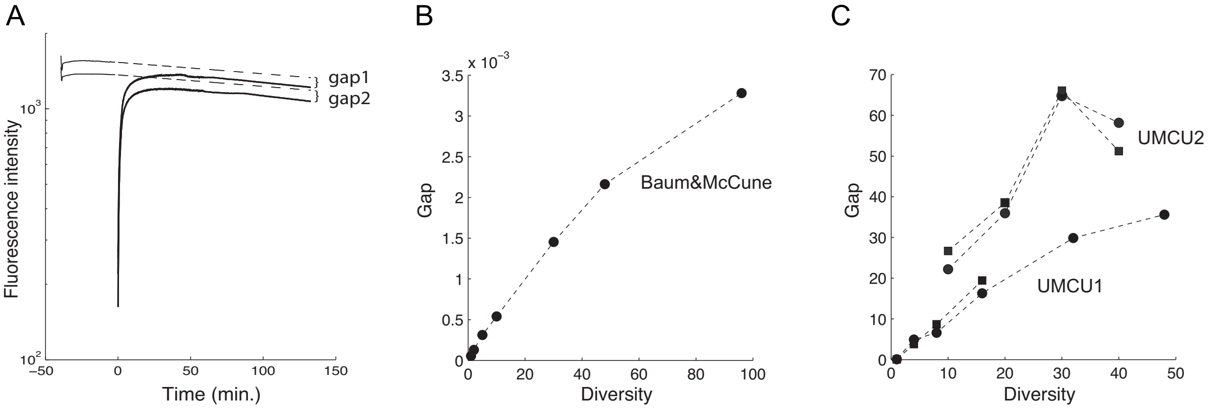


Figure 4.8: Diversity-dependent fluorescence loss. A: The fluorescence gap is defined as the difference between the inferred unmelted sample (dashed lines) and the "limit" measured values of the annealing samples (solid lines) (in this example: $n = 40$ of UMCU data set 2, both replicates). The gap increases with diversity. B: Baum & McCune's data, C: UMCU data sets 1 and 2 (•: replicate 1, ▪: replicate 2). The correlation coefficients and their p-values are in the legends.

The same property was observed to a greater extent in Schütze et al. (2010). The authors of this study constructed samples with very large and known diversities (up to 10^{12} different sequences). When they applied the AmpliCot protocol to their oligonucleotide library, despite the prolonged annealing times (up to 6h.), the fluorescence of samples with diversity 10^6 and above stayed well below 50% of its initial value. The authors were therefore unable to measure the Cot 50% value necessary for applying the Cot-based prediction method. Hence, as suggested by their Figure 2(a), high diversities seem to be associated with a systematic loss of fluorescence.

Contrary to second order kinetics, which implies that a sample should always reach its pre-anneal fluorescence level after re-annealing, the complete model is compatible with the above observations. We hypothesize that the diversity-dependent fluorescence loss is caused by the presence of heteroduplexes, which have a decreased fluorescence intensity.

4.3.8 Fluorescence Intensity for Very Large n

In the previous section, we saw that the fluorescence gap is increasing with diversity. In other terms, the limit fluorescence intensity is inversely correlated with diversity. Here, we investigate the lower bound of the fluorescence, i.e., we study the fluorescence intensity for very large diversities. The question that we want to answer is "Would the samples with very large diversity loose all their fluorescence?" This is of interest, because AmpliCot is designed to estimate the diversity of TCRs that is orders of magnitude larger than the “toy" data sets of our study.

By using the analytical solution of the CM under a quasi-steady state assumption (Section 5.3), we can derive the time-limit fluorescence intensity $F^*(n)$ as a function of diversity:

$$F^*(n) := \lim_{t\to\infty} F(t) = 2(D^*(n) + \varphi J^*(n)),$$

where $D^*(n)$ and $J^*(n)$ are given by Eq. (5.21) and Eq. (5.20) of Chapter 5.

As n increases, the fluorescence level is decreasing and has limit value

$$\tilde{F} := \lim_{n\to\infty} F^*(n) = 1 - (1 - \varphi)\alpha. \tag{4.16}$$

(See also Eq. (5.24) of Chapter 5 for derivation details.) The expression of $\tilde{F}$ (Eq. (4.16)) contains parameters φ and α, i.e., the fluorescence intensity of heteroduplexes and the initial proportion of melted material, respectively. Let us examine the values of these parameters in order to determine the lower bound of the fluorescence intensity for very diverse samples. There are four extreme cases of φ and α, which lead to three distinct values of $\tilde{F}$:

(i) $\varphi = 1$ or $\alpha = 0 \Rightarrow \tilde{F} = 1$

(ii) $\varphi = 0 \Rightarrow \tilde{F} = 1 - \alpha$

(iii) $\alpha = 1 \Rightarrow \tilde{F} = \varphi$

Point (i) is trivial: if heteroduplexes fluoresce as much as homoduplexes ($\varphi = 1$), the CM is reduced to a variant of SOK where the DNA association occurs in two steps (heteroduplexes are not distinguished from homoduplexes) and hence the fluorescence level after re-annealing would reach its pre-anneal initial value ($\tilde{F} = 1$); the case $\alpha = 0$ means that none of the material has melted, which is unrealistic.

Point (ii) means that if heteroduplexes do not fluoresce ($\varphi = 0$), very diverse samples would loose most of their fluorescence and the only remaining fluorescent material, if any, would be the proportion $1 - \alpha$ of unmelted homoduplexes. This can also be seen from the individual contributions of homo- and hetero- duplexes (see Figure 5.2 of Chapter 5). As n increases, the proportion of heteroduplexes reaches the limit value α (Eq. (5.23)), whereas the only remaining homoduplexes are those that have not dissociated during the melting step. Hence, if heteroduplexes do not fluoresce, an "ideal" experiment in which the all the material melts ($\alpha = 1$) would result in the quasi-total loss of fluorescence in very diverse samples ($\tilde{F} \to 0$).

Symmetrically, point (iii) indicates that in an "ideal" experiment where the all the material melts ($\alpha = 1$), there would be only heteroduplexes at the end of the annealing and the maximal fluorescence would be given by their fluorescence ($\tilde{F} = \varphi$).

4.3.9 Diversity Predictions

In order to improve the diversity extrapolations of AmpliCot, we propose here two alternatives to the Cot-based prediction method: the *model-based* and the *t_p-based* methods. Both procedures take advantage of the following results of our modeling:

- The CM is a better description of AmpliCot than SOK;
- Cot scaling does not correct for concentration differences;
- The relation t_p vs n is not linear.

The *model-based method* is based on the first result above. It uses not only the information of the annealing curves sampled at one point as in the Cot-based procedure, but the information contained in the entire annealing curve. It also has the advantage of using directly the raw data, which avoids the introduction of bias and the loss of information due to normalization. The model-based method uses the samples of known diversity to fit the parameters of the underlying annealing kinetics model. The biochemical reaction rates are then fixed to their best-fit values and a range of possible values for the unknown diversity is determined. For each value of n in this pre-determined set, the remaining parameters (Ω_i, Eq. (4.12)) are fitted to the sample with unknown diversity by minimizing the squared errors. The log-likelihood of the resulting fit is calculated. The unknown diversity is estimated as the value of n that maximizes the log-likelihood (see Table 4.2 for a summary of all prediction methods).

The *t_p-based method* is a modified version of the Cot-based one. It uses the second and third results above. Similarly to the Cot-based procedure, it samples the data at a single point (which has to be chosen *a priori*), but instead of assuming a linear relation between t_p and n, it uses the best-fit of the complete model to infer the (possibly nonlinear) relation. The unknown diversity is read from the relation $t_p^{-1}(t)$. The advantage of this method with respect to the model-based one is to ignore the regions of the annealing curve where the model does not fit the data well. The drawback is that one has to choose a “good" t_p value (we discuss this topic in Section 4.3.10).

In order to compare their performance, we have applied the three prediction methods to Baum’s data and to UMCU data set 1. All but the two largest diversity samples were used for calibration ($n = 1, 2, 5, 10, 30$ for Baum’s data, $n = 1, 4, 8, 16$ for UMCU data) and the goal was to predict the largest diversity ($n = 96$ for Baum’s data, $n = 48$ for UMCU data). UMCU data set 2 was left out, because it contained too few diversities for the calibration set. Table 4.4 displays the results of the prediction experiments. The confidence intervals were computed by bootstrapping the calibration sets, as explained in Section 4.2.11.

Method	Data set	Prediction	Error	80% CI
Cot-based (50%)	Baum	71	-26.04%	[71, 75]
	UMCU 1	36	-25.00%	[31, 38]
Cot-based (80%)	Baum	86	-10.42%	–*
	UMCU 1	62	+29.17%	[59, 73]
Model-based	Baum	94	-2.00%	[61, 135]
	UMCU 1	53	+10.00%	[39, 53]
t_p-based (85%)	Baum	97	+1.04%	[73, 129]
	UMCU 1	45	-6.25%	–*

Table 4.4: Prediction results. The three methods are described in Table 4.2. All methods were calibrated on all but the two largest diversity samples. The diversity to predict was $n = 96$ for Baum's data and $n = 48$ for UMCU data. The confidence intervals (CI) were computed by bootstrapping the calibration data sets (see Section 4.2.11). The error is expressed as the percentage from the true diversity. *The CI could not be computed because the distribution of bootstrap replicates exhibited a bi-modal shape (see Figure 4.9).

When comparing the predictions obtained by the Cot-based method with those of the model-based one, we observe an impressive improvement in the case of Baum's data, although this data set was nicely explained with SOK. Indeed, the error in the estimation of the diversity is reduced more than 10 times, passing from 26% to 2% underestimation. The improvement is more moderate in the case of the UMCU data: the error is reduced a bit more than twice, passing from 25% underestimation to 10% overestimation. Using the 80% annealing point instead of 50% in the Cot-based method improves the predictions of Baum's data, but not of the UMCU data. For the t_p-based method, we show here the results with the highest possible annealing percentage (85%). It turns out that this method gives the smallest prediction error for both data sets. The prediction is now within 10% from the true value and the improvement from the initial Cot 50% method is about 5-fold. Thus, the new prediction methods seem to yield good results. This was indeed expected, as these methods use more data information than the Cot-based method.

However, the confidence intervals of the predictions reveal a very large variability around the values predicted by the model-based and t_p-based methods. Nevertheless, these confidence intervals contain the true diversity, which is not the case for the Cot-based method. To further analyze the performance of the prediction methods, we have plotted the histograms of bootstrapped predictions (Figure 4.9). Each prediction method was applied to R bootstrap replicates and the histogram of resulting diversity estimations is shown (these distributions were used for the confidence intervals estimations). The true diversity values are indicated by a dashed vertical line. We note that the density of predictions of the Cot-based method (both for $p = 0.5$ and

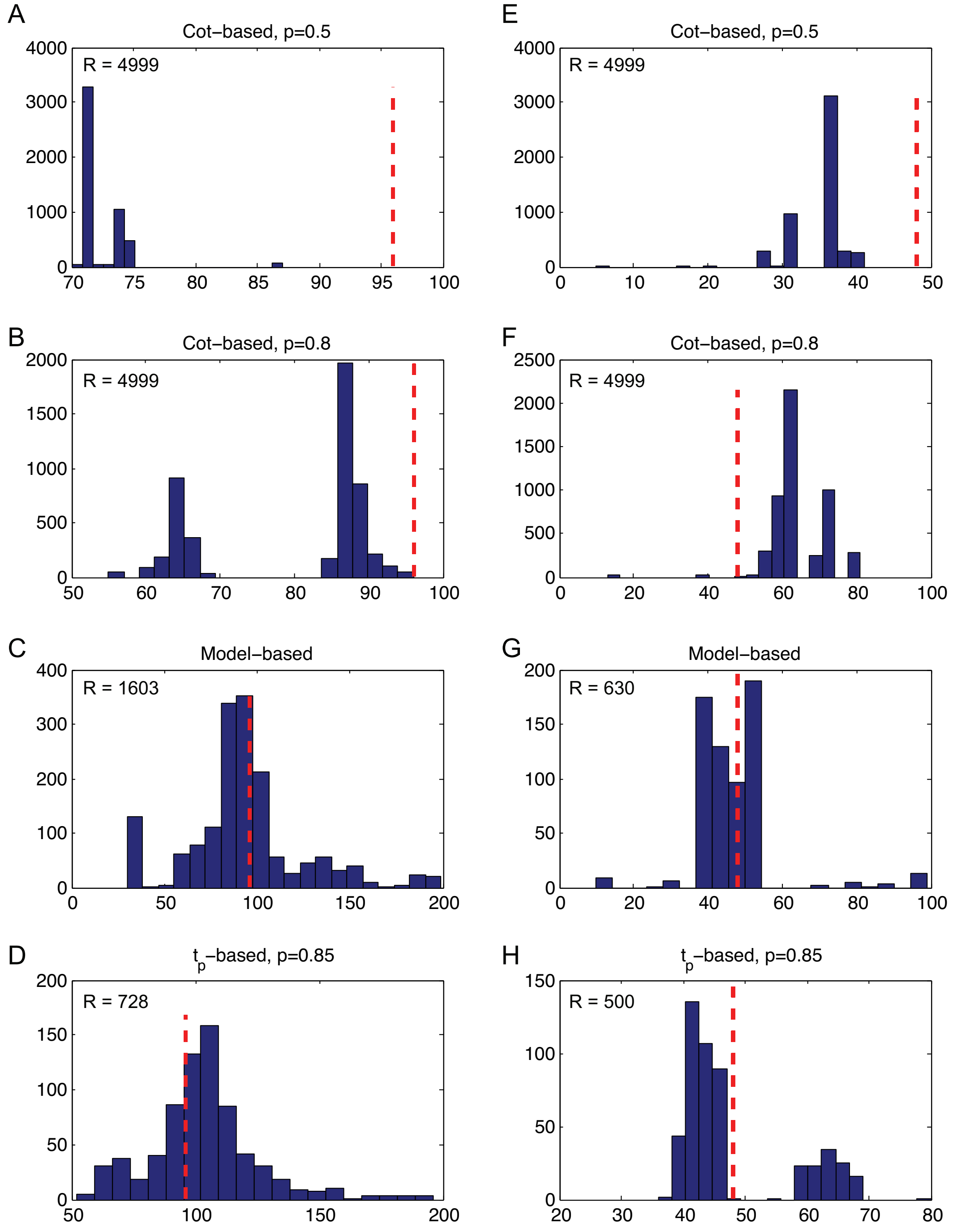


Figure 4.9: Bootstrap distributions. The Cot-based (first and second row of panels), model-based (third row of panels) and the t_p-based method (last row of panels) were executed on bootstrap replicates of the calibration sets of Baum's data set (panels A-D) and UMCU data set 1 (panels E-F). The histograms of the resulting predictions are plotted. The search diversity is indicated by the red dashed vertical line. The effective number of bootstrap replicates R is indicated in the upper left corner of each plot. Although the new methods are very sensitive, their distribution is centered around the true diversity.

0.8, panels A,B,E,F) are far from the true diversity. Remark moreover the bimodal distribution of panel B, which indicates that the Cot-based method applied on Baum's data can yield a reasonable diversity prediction. The model-based and the t_p-based methods have densities centered around the true value. Again, a bimodal distribution is revealed by the t_p-based method applied on UMCU data set 1 (panel H), which prevented us from computing a confidence interval. Note that predicting a diversity from a bootstrap replicate is rather challenging because with high probability, the calibration set contains less diversities than the original calibration set. This can explain why the methods that involve more parameters (the model- and t_p-based) exhibit distributions with larger variance.

4.3.10 Criteria for Choosing a Correct Annealing Percentage

The Cot- and the t_p-based prediction methods have one parameter, the annealing percentage p, that needs to be set before applying the procedure. A last question remains: how to choose an annealing percentage such that the prediction errors are minimized? Here, we investigate the dependence of the prediction on the chosen value of p. Figure 4.10 shows this dependence for both data sets (Baum: A, UMCU: B) and for both prediction methods. Note that the Cot-based method assumes second order kinetics, whereas the t_p-based method is based on the complete model.

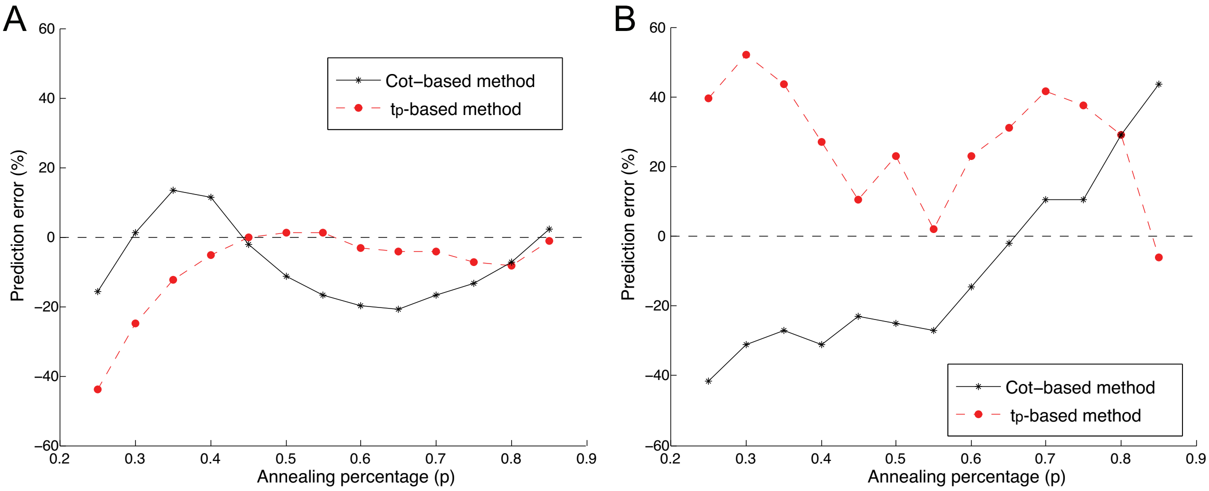


Figure 4.10: Prediction error as function of the annealing percentage. The Cot- (resp. t_p-) based method was executed with different values of p and the prediction error was computed. A: Baum's data, calibration on $n = 1, 2, 5, 10, 30$, predicting $n = 96$, B: UMCU data set 1, calibration on $n = 1, 4, 8, 16$, predicting $n = 48$.

For Baum's data, the t_p-based method is in general producing a smaller error than the Cot-based version (except for values of p below 40%). In general, for both methods, it seems that the prediction is improving as the annealing percentage is increasing. In fact, both methods give similar predictions for Cot values above 80%. This is not the case for UMCU data (Figure 4.10B). For this data set, the t_p-based method has a more "chaotic" behavior, but it still seems that higher annealing percentages result in better predictions.

We searched for criteria that would guide the choice of annealing percentage from the calibration set, before making a diversity extrapolation. In other words, we aimed to find the region of the data that contains most information about the diversity and that can be inferred from the calibration data set.

For the t_p-based method, a possible candidate would be the quality of the fit. As we saw in Figure 4.2, the complete model fits some parts of the annealing curve better than others. For example, the first 10 minutes of Baum's experiment are not very well explained with our model, thus we would like to avoid the annealing percentages corresponding to that region. To see whether there is a positive correlation between the sum of squared errors (SSE) around an annealing percentage p and the prediction error, we have plotted the latter quantities (Figure 4.11). There is clearly a positive correlation between both variables (corr. coeff. = 0.83 for Baum's data, 0.67 for UMCU data), thus confirming that in general, a good fit implies a good prediction. However, the relation is not really linear (especially for Baum's data) and very good predictions can be obtained with not-so-good fits. Nevertheless, if we use the annealing point at which the SSE is minimal (p = 0.85 for Baum's data, and p = 0.47 for UMCU data, figure not shown), we obtain a reasonably low error in the estimate of the diversity (+1.05% for Baum and +12.5% for UMCU).

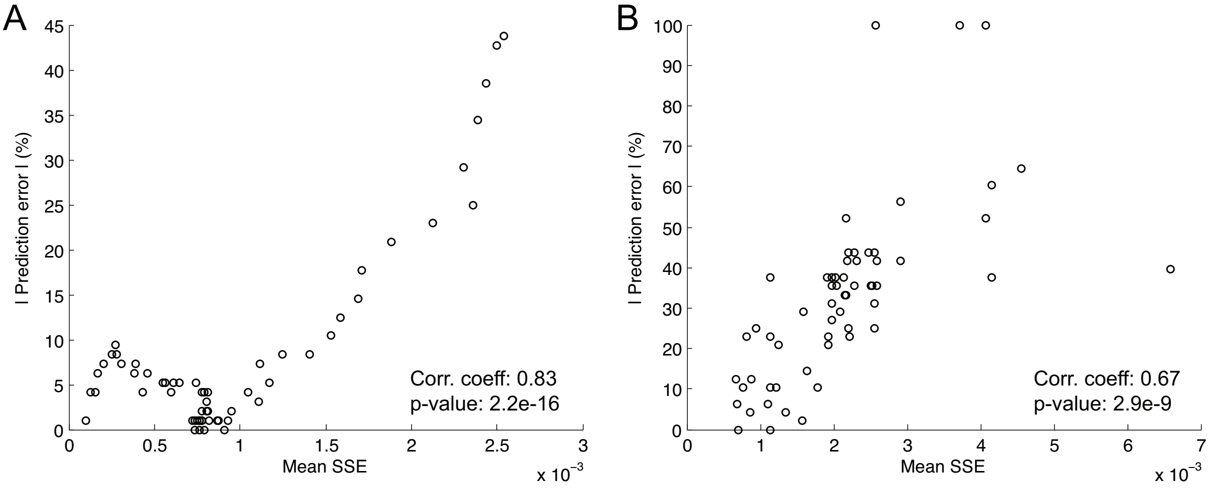


Figure 4.11: Mean sum of squared errors (SSE) vs prediction error. The t_p-based method was executed with different values of p. The quality of fit at the value of p, in terms of the mean SSE, is plotted horizontally; the absolute value of the relative prediction error is plotted vertically. A: Baum's data, B: UMCU data set 1. A significant positive correlation is observed.

For the Cot-based method, a possible rule could be the quality of the linear regression (in terms of the R^2 statistic). However, we did not find a positive correlation between the latter and the prediction error. Therefore, we recommend to try the Cot-based method with several Cot values in order to test the sensitivity of the technique to these values. A higher Cot value can give better results, but according to our analysis, it is not guaranteed.

4.4 Discussion

By means of a mathematical model, we analyzed and improved the interpretation of AmpliCot (Baum and McCune, 2006). We showed that the initially assumed underlying model, second order kinetics, is not always the best to describe the annealing dynamics. We propose an alternative, the complete model, which considers the biochemical reaction in further detail. The consequences of using this alternative model are twofold: the Cot scaling that was applied under SOK to account for concentration differences between samples is no longer applicable, and the relation between t_p values and diversity is no longer linear. These facts have casted doubt on the accuracy of the diversity extrapolation approach based on second order kinetics. As a solution, we have proposed two alternative prediction methods that take into account the findings of our modeling.

Our model-based prediction method has several advantages over the Cot-based one. It is independent on a single annealing percentage, because it considers the entire annealing curve in order to extrapolate the necessary information. Thus, it uses the full data set, not only a subset of it. It also yielded good prediction results when applied to the extrapolation of a sample with diversity at least 3 times larger than the diversities of the calibration set.

A possible drawback of our approach would be its sensitivity to the correct estimation of the bio-chemical reaction rates. As we saw, the confidence limits of the parameter estimates are sometimes large (Table 4.3), suggesting that some parameters might not be separately identifiable. A solution to that problem could be to fit directly the combined parameters $\frac{d_1}{(d_1+z_1)}$ and $\frac{d_2}{(d_2+z_2)}$.

Another source of inaccuracy might appear on the level of the “diversity fitting" itself. Once the best-fit reaction rates are found, the diversity extrapolation procedure consists in iterating over a pre-determined range of diversities $n \in \bar{N}$ and fitting the remaining curve-specific parameters (see Table 4.2 for a summary of the procedure). The value of n that maximizes the likelihood of the unknown diversity sample is chosen as the searched diversity n^*. A legitimate question is whether the optimal value n^* is a global maximum of the likelihood function? In other words, can we identify n given the best-fit values of the reaction rates? To address this issue, we have artificially generated data according to both models (SOK and the CM) by adding random noise to the numeric solution of the ODEs of Eq. (4.3) or Eq. (4.5) (data not shown). The artificial data was then used to perform the prediction experiments in the same way it was performed with the real data. When the same model was used to generate and to fit the artificial data, the resulting predictions were 100% accurate. This indicates that once the parameter values are found by the fitting procedure, there is an unique value of n for which the likelihood of the data set with unknown diversity is maximal. Note that this is true for an uniform species distribution, but it might not be true in presence of expanded clones.

A parameter that strongly influences the output of the model is the baseline fluorescence of a sample (A_b). This parameter determines the post-annealing fluorescence level, which, in turn, determines the value of φ, the fluorescence intensity of heteroduplexes (and maybe even the reaction rates determining the fraction of heteroduplexes). In order to obtain the best possible estimates of A_b, our approach requires long enough pre-anneal measurements. This was not the case with Baum & McCune’s raw data, where the pre-anneal values lasted for only 10 minutes. The problem with such short measurements is the possible instability of the fluorescence intensity, as seen in this data set. A clear example of this instability can be seen in the sample of diversity $n = 1$ (Figure 4.2). Indeed, the fluorescence intensity clearly increases during the pre-anneal measurements, whereas a slow decline is expected due to dye degradation. These transient instabilities are influencing the estimation of A_b and thus adding noise to the rest of the parameter estimates. Good estimates of A_b are needed as well for the normalization transformation (Eq. (4.13)), hence we recommend that pre-anneal measurements should last long enough (at least 20-30 minutes).

Another issue related to the estimation of A_b, but also to other parameters, is the time interval between the true beginning of the annealing and the first measurement (we call this time interval Δ). We assume that we know the initial conditions of the ODE system at the beginning of the annealing phase, but the first available measurement is some time after the start of the annealing. If this time lapse (Δ) is large, it might influence the estimation of the reaction rates. Parameter Δ also influences the value of the limit fluorescence φ, because A_b, which is determined by the last pre-anneal measurements, is corrected for the time-dependent fluorescence decay between the last pre-anneal point and the first annealing measurement. It is therefore important to have a precise knowledge of this time gap. In order to obtain the best-possible parameter estimates, we tried adding Δ to both models as an unknown parameter to be fitted. However, as the resulting best-fit values of Δ turned out to be small (less than a minute for all data sets), for simplification, we have removed this parameter from our main model description.

The complete model was able to explain the fact that the re-annealed samples do not reach their pre-anneal fluorescence intensity, even after correction for the fluorescence decline due to dye degradation, as was noted in Schütze et al. (2010). We suggest that this is due to the presence of heteroduplexes which have a lower fluorescence than homoduplexes. An alternative explanation of this phenomena would be that as the annealing temperature is very close to the melting one, the formed homoduplexes dissociate and re-associate again. We have fitted such a model to the data and although it accounted for the above-mentioned fluorescence gap, this model did not explain the beginning of the annealing curves and therefore yielded a significantly lower quality of the fit with respect to the complete model. An alternative explanation of the fluorescence gap could be a diminished efficiency of the SYBR green die after melting. However, this would not explain the diversity-dependence of the gap.

An important issue that was not addressed at this stage is the effect of expanded clones. We studied here only the case of an uniform distribution of species, which allowed us to reduce the dimension of the ODE systems. This resulted in a computationally efficient definition of the models, independent of the number of species. The equimolarity assumption is probably valid for naive T cells, but its validity is lost in the case of antigen-primed T cell populations (Naumov et al., 2003), although it was recently shown that the human memory repertoire is composed of a majority of low frequency clones (Klarenbeek et al., 2010). Nevertheless, the use of AmpliCot in the presence of even one expanded clone becomes more tricky. If a sample contains one or several expanded clones, the annealing kinetics of this sample would behave as if its diversity is lower. Fitting the complete model might become very complex in that situation and further analysis is needed in order to make a conclusion. In such cases, reading the annealing curves at a single point as close to completion as possible, as suggested in Baum and McCune (2006), would be the best solution at this moment.

Finally, in addition to the development of methods dealing with non-equimolar samples, it would be of interest to validate our current results with samples of diversity higher than 100 species.

Chapter 5

Analysis of the AmpliCot Model

In Chapter 4, we defined two models describing the annealing kinetics of AmpliCot, namely second order kinetics (SOK) and the complete model (CM) (Section 4.2.3). In this chapter, we derive the theoretical results necessary for the analysis of these models. In Section 5.1, we formally expose the correct way of handling concentration differences under the CM. We then rigorously show why Cot scaling is not valid under the CM (Section 5.2). Finally, we present in Section 5.3 an analytic solution of the CM, valid in the particular case of a quasi-steady state assumption (QSS).

5.1 Handling Concentration Differences

The theorem exposed in this section shows that comparing samples with different concentrations affects the association rates. In fact, correcting the association rates is an alternative to Cot scaling in accounting for the concentration differences between samples. The advantage is that this correction can be used with both models, not only with second order kinetics.

Before giving the theorem, we introduce a general framework for describing the models. The differential equations can be written as:

$$\begin{aligned} \frac{d\mathbf{x}(t;\Theta,T)}{dt} &= F(\mathbf{x},\Theta), \\ \mathbf{x}(0;\Theta,T) &= T * (\mathbf{f},\vec{0}_{m-n})' \end{aligned} \tag{5.1}$$

where $\mathbf{f} = (f_1, f_2, \dots, f_n)$ is the vector giving the proportion of each DNA type inside the sample ($\sum_{i=1}^{n} f_i = 1$), $\vec{0}_{m-n} = (0, 0, \dots, 0)$ is the zero vector of dimension $1 \times (m-n)$, m is the number of state variables, n is the sample diversity, and $\mathbf{x}$, Θ and $F(\cdot)$ are defined differently according to the model.

We split the parameters in two groups, $\Theta := (\theta_1, \theta_2)$, in order to distinguish the parameters that have units depending on the concentration ($\theta_1 = a$) from those having time-dependent units ($\theta_2 = (d_1, d_2, z_1, z_2)$).

Theorem 2. *The normalized dynamics of a sample with total concentration of ssDNA* $S_0 = \gamma \neq 1$ *and reaction rates* $\Theta = (\theta_1, \theta_2)$ *are equivalent to the dynamics of a sample with total* *concentration of ssDNA* $S_0 = 1$ *and reaction rates* $\Theta = (\gamma\theta_1, \theta_2)$. *Mathematically,*

$$\frac{\mathbf{x}(t;\theta_1,\theta_2,\gamma)}{\gamma} = \mathbf{x}(t;\gamma\theta_1,\theta_2,1).$$

Proof. Let

$$\mathbf{y}(t;\theta_1,\theta_2,\gamma) := \gamma\mathbf{x}(t;\gamma\theta_1,\theta_2,1). \tag{5.2}$$

We will show that $\mathbf{y}$ satisfies the same differential equation as $\mathbf{x}(t;\theta_1,\theta_2,\gamma)$, which is

$$\begin{aligned} \frac{d}{dt}\mathbf{x}(t;\theta_1,\theta_2,\gamma) &= F(\mathbf{x}(t;\theta_1,\theta_2,\gamma),\theta_1,\theta_2) && (5.3)\\ \mathbf{x}(0;\theta_1,\theta_2,\gamma) &= \gamma(\mathbf{f},\vec{0}_{m-n}) && (5.4) \end{aligned}$$

For the initial conditions, we have by definition:

$$\begin{aligned} \mathbf{y}(0;\theta_1,\theta_2,\gamma) &= \gamma\mathbf{x}(0;\gamma\theta_1,\theta_2,1)\\ &= \gamma(\mathbf{f},\vec{0}_{m-n})'\\ &= \mathbf{x}(0;\theta_1,\theta_2,\gamma). \end{aligned}$$

Then, using the definition of $\mathbf{y}(\cdot)$,

$$\begin{aligned} \frac{d}{dt}\mathbf{y}(t;\theta_1,\theta_2,\gamma) &= \gamma\frac{d}{dt}\mathbf{x}(t;\gamma\theta_1,\theta_2,1)\\ &= \gamma F(\mathbf{x}(t;\gamma\theta_1,\theta_2,1),\gamma\theta_1,\theta_2)\\ &\overset{\text{Eq. (5.2)}}{=} \gamma F\left(\frac{1}{\gamma}\mathbf{y}(t;\theta_1,\theta_2,\gamma),\gamma\theta_1,\theta_2\right)\\ &= F(\mathbf{y}(t;\theta_1,\theta_2,\gamma),\theta_1,\theta_2), \end{aligned}$$

where the last equality holds for both models. Indeed, we show below that it is true for the function F in the case of the complete model and therefore, it is also true for SOK, as it is a subset of the former. Let $F(\cdot) = (F_1(\cdot),\ldots,F_1(\cdot))$, where $F_1(\cdot),\ldots,F_1(\cdot)$ are the right-hand sides of the system of Eq. (4.2). We have

$$\begin{aligned} \gamma F_1\left(\frac{1}{\gamma}\mathbf{y}(t;\theta_1,\theta_2,\gamma),\gamma\theta_1,\theta_2\right) &= \gamma\left(-2a\gamma\frac{S_i^2}{\gamma^2} - a\gamma\frac{S_i}{\gamma}\sum_{j\neq i}\frac{S_j}{\gamma} + 2d_1\frac{C_{ii}}{\gamma} + d_2\sum_{j\neq i}\frac{C_{ij}}{\gamma}\right)\\ &= -2aS_i^2 - aS_i\sum_{j\neq i}S_j + 2d_1C_{ii} + d_2\sum_{j\neq i}C_{ij}\\ &= F_1(\mathbf{y}(t;\theta_1,\theta_2,\gamma),\theta_1,\theta_2) \end{aligned}$$

and similarly for F_2, F_3 and F_4.

The function $F(\mathbf{y},\Theta)$ being continuous in $\mathbf{y}$ and having continuous first partial derivatives with respect to $\mathbf{y}$, there is a unique solution to the initial value problem of Eq. (5.3). Therefore, as $\mathbf{y}(\cdot)$ is a solution of the differential equation Eq. (5.3), we must have $\mathbf{y}(\cdot) \equiv \mathbf{x}(\cdot)$, i.e., $\mathbf{x}(t;\theta_1,\theta_2,\gamma) = \gamma\mathbf{x}(t;\gamma\theta_1,\theta_2,1)$, which proves the theorem. □

Theorem 2 means that normalized data should not be fitted with unique association rates θ_1 for all diversity samples, but sample i should have association rate $\gamma_i\theta_1$, where γ_i is an estimate of the total concentration of sample i.

5.2 Cot Scaling is Only Valid under Second Order Kinetics

Consider the differential equations defining second order kinetics:

$$\begin{aligned} \frac{dS_i}{dt} &= -2aS_i^2 \\ \frac{dD_{ii}}{dt} &= aS_i^2, \end{aligned} \tag{5.5}$$

$i = 1, \ldots, n$. In compact form, these can be written as

$$\frac{d\mathbf{x}}{dt} = F(\mathbf{x}(t)), \tag{5.6}$$

where $\mathbf{x} = (S_1, \ldots, S_n, D_{11}, \ldots, D_{nn})'$ is a solution of Eq. (5.6). Let

$$\mathbf{y}(t) := T\mathbf{x}(Tt), \tag{5.7}$$

i.e., a Cot re-scaled system with total concentration T. Differentiating Eq. (5.7) with respect to t yields

$$\frac{d\mathbf{y}}{dt} = T^2 \frac{d\mathbf{x}}{dt}(Tt) = T^2 F(\mathbf{x}(Tt)) \tag{5.8}$$

In the case of second order kinetics (Eq. (5.5)), F is a homogeneous function, i.e., $F(Tt) = T^2 F(Tt)$. Indeed, we have

$$F\begin{pmatrix} S_1(t) \\ \vdots \\ S_n(t) \\ D_{11}(t) \\ \vdots \\ D_{nn}(t) \end{pmatrix} = \begin{pmatrix} -2aS_1(t)^2 \\ \vdots \\ -2aS_n(t)^2 \\ aS_1(t)^2 \\ \vdots \\ aS_n(t)^2 \end{pmatrix}$$

and

$$F\begin{pmatrix} TS_1(Tt) \\ \vdots \\ TS_n(Tt) \\ TD_{11}(Tt) \\ \vdots \\ TD_{nn}(Tt) \end{pmatrix} = \begin{pmatrix} -2aT^2S_1(t)^2 \\ \vdots \\ -2aT^2S_n(t)^2 \\ aT^2S_1(t)^2 \\ \vdots \\ aT^2S_n(t)^2 \end{pmatrix} = T^2 F(\mathbf{x}(Tt)).$$

Therefore,

$$\frac{d\mathbf{y}}{dt} = T^2 F(\mathbf{x}(Tt)) = F(T\mathbf{x}(Tt)) = F(\mathbf{y}(t)),$$

meaning that $\mathbf{y}(t)$ is also a solution of Eq. (5.6). Thus, Cot scaling is correcting for concentration differences under second order kinetics.

Cot scaling is not valid for the complete model, because the function F defining the system is not homogeneous.

5.3 Analysis of the CM under a QSS Assumption

We present an analytical solution of the complete model in the particular case of a quasi-steady state (QSS) assumption. Such an assumption is valid if the association and dissociation rates (a, d_1, d_2) of transient complexes are large compared to the polymerization rates z_1 and z_2 (alternatively, if there are two time scales in the dynamics (Borghans et al., 1999)). By applying the QSS assumption, we can solve the ODE system (Section 5.3.1) and find equilibrium points (Section 5.3.2). The solution is further used to the fluorescence intensity for large diversity samples (Section 5.3.3) and to derive an analytical expression of $t_p(n)$ (Section 5.3.4).

5.3.1 Complete Model Solution

The following system of differential equations defines the complete model (assuming equimolar distribution of species):

$$\frac{dS}{dt} = -a(n+1)\frac{S^2}{n} + 2d_1C + 2d_2H \tag{5.9}$$

$$\frac{dC}{dt} = a\frac{S^2}{n} - (d_1 + z_1)C \tag{5.10}$$

$$\frac{dH}{dt} = a\left(\frac{n-1}{2}\right)\frac{S^2}{n} - (d_2 + z_2)H \tag{5.11}$$

$$\frac{dJ}{dt} = z_2H \tag{5.12}$$

$$\frac{dD}{dt} = z_1C, \tag{5.13}$$

with initial conditions $S(0) = \alpha T$, $2D(0) = (1-\alpha)T$, $C(0) = H(0) = J(0) = 0$, and conservation law

$$S(t) + 2(C(t) + H(t) + D(t) + J(t)) = T. \tag{5.14}$$

If the association/dissociation rates a, d_1, d_2 of transient complexes $C(t)$ and $H(t)$ are large compare to the final duplex formation rates z_1 and z_2, we can assume that the transient complexes reach quickly a steady-state. By setting their corresponding time-derivatives (Eq. (5.10) and Eq. (5.11)) to 0, we get:

$$C = \frac{a}{n(d_1 + z_1)}S^2 \tag{5.15}$$

$$H = \frac{a}{n(d_2 + z_2)}\left(\frac{n-1}{2}\right)S^2 \tag{5.16}$$

By inserting Eq. (5.15) and Eq. (5.16) in the initial system, we get:

$$\begin{aligned} \frac{dS}{dt} &= -2K(n)S^2 \\ \frac{dJ}{dt} &= \frac{a}{n}\left(\frac{z_2}{d_2 + z_2}\right)\left(\frac{n-1}{2}\right)S^2 \\ \frac{dD}{dt} &= \frac{a}{n}\left(\frac{z_1}{d_1 + z_1}\right)S^2, \end{aligned} \tag{5.17}$$

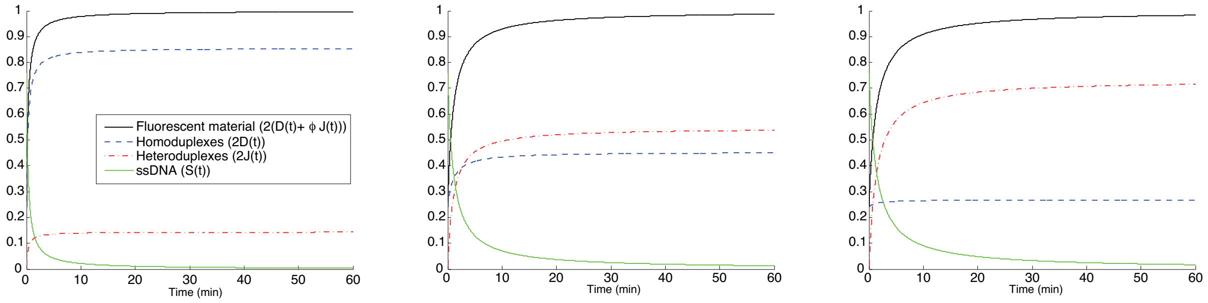
where

$$K(n) = \frac{a}{n}\left(\frac{z_1}{d_1+z_1} + \frac{z_2}{d_2+z_2}\left(\frac{n-1}{2}\right)\right). \tag{5.18}$$

By using the initial conditions, we obtain the following solution of Eq. (5.17):

$$\begin{aligned} S(t;n) &= \frac{\alpha T}{1+2K(n)\alpha T t} \\ J(t;n) &= \frac{a}{n}\left(\frac{z_2}{d_2+z_2}\right)\left(\frac{n-1}{2}\right)\frac{1}{2K(n)}(\alpha T - S(t)) \\ D(t;n) &= \frac{a}{n}\left(\frac{z_1}{d_1+z_1}\right)\frac{1}{2K(n)}(\alpha T - S(t)) + \left(\frac{1-\alpha}{2}\right)T, \end{aligned} \tag{5.19}$$

where we have stressed the dependance on the diversity by putting n in argument.



(a) $n = 10$ (b) $n = 100$ (c) $n = 1000$

Figure 5.1: Dynamics of the complete model under a quasi-steady state assumption for three different diversities. Eq. (5.19) has been simulated with parameter values from Table 4.3, UMCU data set 1, total concentration $T = 1$ in all panels. Fluorescent material (cf. Eq. (5.22)): solid black. Proportion of heteroduplexes: dashed blue. Proportion of homoduplexes: dot-dashed red. ssDNA: solid green.

To gain some insight of the above dynamics, Figure 5.1 depicts the three functions of Eq. (5.19) with parameter values from the best-fit of UMCU data set 1 (Table 4.3) and three different values of n ($n = 10, 100, 1000$). Note that the total ssDNA concentration was set to $T = 1$ for each n. As diversity increases, the overall annealing speed, given by the sum $2(D(t;n) + \varphi J(t;n))$, is slowing down. Moreover, the proportion of heteroduplexes increases with diversity, whereas the proportion of homoduplexes decreases. This is intuitive since the probability for a single-stranded DNA to find its perfectly matching mate in the same number of encounters is smaller when the total number of species is large.

5.3.2 Time Limit

An equilibrium point of the system under the quasi-steady state assumption is obtained by taking the time limit in Eq. (5.19):

$$\begin{aligned} S^*(n) &:= \lim_{t\to\infty} S(t;n) = 0 \\ J^*(n) &:= \lim_{t\to\infty} J(t;n) = \frac{a}{n}\left(\frac{z_2}{d_2+z_2}\right)\left(\frac{n-1}{2}\right)\frac{\alpha T}{2K(n)} \qquad (5.20) \\ D^*(n) &:= \lim_{t\to\infty} D(t;n) = \left(\frac{a}{n}\left(\frac{z_1}{d_1+z_1}\right)\frac{\alpha}{2K(n)}+\frac{1-\alpha}{2}\right)T, \qquad (5.21) \end{aligned}$$

where $K(n)$ is defined by Eq. (5.18). The limit values of C and H are 0. The time-limit of the fluorescence intensity is then

$$F^*(n) := \lim_{t\to\infty} F(t) = 2(D^*(n) + \varphi J^*(n)). \qquad (5.22)$$

These limit values are illustrated in Figure 5.1.

5.3.3 Diversity Limit

We now consider the equilibrium points of Eq. (5.20) and Eq. (5.21). We want to compute the limit $n \to \infty$ of $J^*(n)$ and $D^*(n)$. By replacing the expression of $K(n)$ (Eq. (5.18)) back in Eq. (5.20) and Eq. (5.21), we get after some algebraic manipulations:

$$\begin{aligned} \lim_{n\to\infty} 2J^*(n) &= \alpha T \qquad (5.23) \\ \lim_{n\to\infty} 2D^*(n) &= (1-\alpha)T. \end{aligned}$$

Interestingly, Eq. (5.23) means that for very large diversities, the only material that is found in homoduplex form is the one that did not initially melt ($(1-\alpha)T$). In other words, if n is large enough, all the single strands that re-anneal would form heteroduplexes.

From Eq. (5.23), we can write the expression of the fluorescence intensity for large diversities:

$$\begin{aligned} \lim_{n\to\infty} F^*(n) &= 2(\lim_{n\to\infty} D^*(n) + \varphi \lim_{n\to\infty} J^*(n)) \\ &= (1-(1-\varphi)\alpha)T \qquad (5.24) \end{aligned}$$

The dynamics of $2D^*(n)$, $2J^*(n)$ and $F^*(n)$ with the best-fit parameters of UMCU data set 1 are shown in Figure 5.2.

5.3.4 Analytical Expression of $t_p(n)$

By using the normalization transformation of Eq. (4.14) and the analytical solution Eq. (5.19) of the CM under the QSS assumption, we can write the expression of the annealing kinetics:

$$\begin{aligned} A(t) &= 2(D(t;\alpha\theta_1,\theta_2,1) + \varphi J(t;\alpha\theta_1,\theta_2,1)) \\ &= \frac{a}{nK(n)}\left(\frac{z_1}{d_1+z_1} + \varphi\frac{z_2}{d_2+z_2}\left(\frac{n-1}{2}\right)\right)(1-S(t;\alpha\theta_1,\theta_2,1)). \end{aligned}$$

We are now searching for the time t_p at which a proportion p of the sample has annealed, i.e., $A(t_p) = p$. After some algebraic manipulations, we find the following expression of t_p as function of n ($n \geq 1$):

$$t_p(n) = \frac{pn}{2\alpha a\left(\left(\frac{z_1}{d_1+z_1}\right)(1-p) + \left(\frac{z_2}{d_2+z_2}\right)\left(\frac{n-1}{2}\right)(\varphi - p)\right)}.$$

The above expression can be written as

$$t_p(n) = \frac{pn}{K_1 + K_2(n-1)}, \tag{5.25}$$

where

$$K_1 = 2\alpha a\left(\frac{z_1}{d_1+z_1}\right)(1-p), \quad K_1 > 0 \tag{5.26}$$

$$K_2 = \alpha a\left(\frac{z_2}{d_2+z_2}\right)(\varphi - p), \quad K_2 \in \mathbb{R}. \tag{5.27}$$

It is valid for $n \neq 1 - K_1/K_2$. Note that for $p = \varphi$, $t_p(n) = pn/K_1$ is a linear function of n. In order to guarantee the monotonicity of $t_p(n)$, p should be smaller than φ, which prevents from using very large annealing percentages. Finally, note that $t_p(n)$ is bounded above ($\lim_{n\to\infty} t_p(n) = p/K_2$), which might be a problem when applying the modified Cot-based prediction method to large diversities. When the QSS assumption is valid, Eq. (5.25) can be inverted and used in the modified Cot-based prediction method for diversity extrapolations.

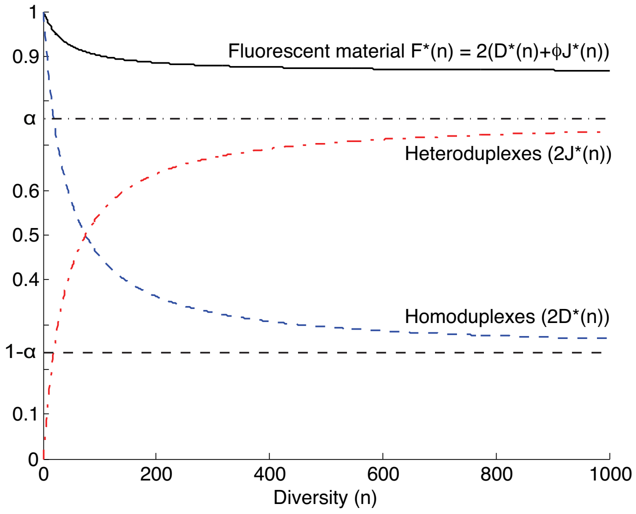


Figure 5.2: The maximal fluorescence level (solid line) is bounded below by the fluorescence of heteroduplexes (here: $\varphi = 0.82$). The proportions of homo- and hetero-duplexes (resp. dashed and dotted-dashed lines) as function of n. Parameter values: best-fit parameters of UMCU data set 1 (see Table 4.3), $\alpha = 0.76$ and $T = 1$.

Concluding Remarks

Chapter 6

Concluding Remarks

In this thesis, we have addressed through mathematical modeling two immunological questions: the lifelong dynamics of regulatory T cells and the measurement of T cell receptor diversity. Our approach was embedded into experimental reality and the developed models were fitted to biological data.

6.1 Lifelong dynamics of regulatory T cells

6.1.1 The questions we addressed

Using a mathematical model, we studied the effect of different sources and division capacities on the lifelong dynamics of the *in vivo* pool of $\mathrm{T_{regs}}$. We identified two properties of the dynamics: a common equilibrium mechanism operating over precursor and mature $\mathrm{T_{regs}}$ populations and a ratio inversion in the early years of adulthood. We identified several biological scenarios that reproduce the above properties. Studying the lifelong *in vivo* dynamics of regulatory T cells is an essential step in the elaboration of a tool able to predict $\mathrm{T_{regs}}$ population size as a function of time. This tool could be used in the fine-tuning of future treatments of pathologies such as transplanted organ rejection or auto-immune diseases.

6.1.2 Recent Advances and Open Questions

Recent experimental investigations have confirmed some of our findings and have further suggested possible modifications to our $\mathrm{T_{regs}}$ model. Here, we cover some of the latest studies that are related to ours.

In our $\mathrm{T_{regs}}$ model, we considered structural subsets of $\mathrm{T_{regs}}$. Precursor $\mathrm{T_{regs}}$ were first identified in 2005 (Fritzsching et al. (2006); Seddiki et al. (2006b); Valmori et al. (2005) and unpublished data by L. Codarri), shortly before the conception of our model. However, their proliferation and suppression capacity was not unanimous at that time, and the differentiation of precursors into mature $\mathrm{T_{regs}}$ was observed only *in vitro*. Furthermore, the phenotypical distinction between activated (R) and quiescent (Q) mature $\mathrm{T_{regs}}$ was not evidenced experimentally. This is why in our model, we considered the dynamics of the sum of both populations.

Recently, Miyara et al. (2009b) have performed a longitudinal study, experimentally addressing part of the questions addressed in our model. Based on the expression level of FoxP3 and CD45RO, Miyara et al. (2009b) identified three phenotypically and functionally distinct subpopulations. The first one, which the authors characterize as CD45RO^{-}FoxP3$^{\text{low}}$, are resting T$_{\text{regs}}$ that have “never" experienced antigen, as suggests the name “naive" they are given in Sakaguchi et al. (2010). These cells correspond to the “precursors" of our model. Miyara et al. (2009b) have confirmed the *in vivo* differentiation pathway suggested in our scheme (and previously observed *in vitro*), in which precursors start proliferating upon TCR stimulation and convert into the mature phenotype, thus expressing CD45RO. Hence, the human *in vivo* transformation from compartments P to R, together with the proliferation capacities of both of these populations, has been experimentally confirmed. However, it seems that the conversion from R to Q and inversely, i.e., the transitions between activated mature T$_{\text{regs}}$ and resting mature T$_{\text{regs}}$, is still an open question (see Sakaguchi et al. (2010) and the “Cell Fate" pictogram of Battaglia and Roncarolo (2009)).

Miyara et al. (2009b) further confirmed the inversion of the ratio precursor/mature T$_{\text{regs}}$, although these authors didn’t study a detailed time-course of both cell sub-populations and therefore could not make a quantitative statement about the age at which the ratio is inverted.

There is a difference, on the level of mature T$_{\text{regs}}$, between our cell sorting and the one of Miyara et al. (2009b). In our case, we have considered mature T$_{\text{regs}}$ as being CD45RO^{+}CD25^{+}FoxP3^{+}. On the other hand, Miyara et al. (2009b) have found that inside CD45RO^{+}CD25^{+}FoxP3^{+} cells, two populations can be further distinguished: a suppressive and proliferative population of CD45RO^{+}FoxP3$^{\text{hi}}$ cells (which corresponds to our activated mature T$_{\text{regs}}$) and a non-regulatory CD45RO^{+}FoxP3$^{\text{low}}$ population. This therefore suggests that our mature T$_{\text{regs}}$ staining may contain some non-regulatory (or unstable regulatory) T cells.

An aspect that was not included in our model is the recently discovered suppressive action of activated mature T$_{\text{regs}}$ on precursors’ proliferation (Miyara et al., 2009b). As a first approximation to the suggested negative feedback loop, one could assume that stimulated precursors do not proliferate as such, but do so once they have acquired the mature phenotype. This scenario was in fact considered among our possible scenarios (proliferation scenario 4) and it was able to explain the data given some parameter constraints. However, a more detailed interaction between these T$_{\text{regs}}$ subsets, in particular the proliferation of precursors dependent on the number of activated mature T$_{\text{regs}}$ of the same clone, could be an interesting future research line. Furthermore, concerning cell-interactions, it would be interesting to include in our model the dependence of the antigen-induced proliferation of T$_{\text{regs}}$ on conventional (non-regulatory) T cells, as suggested by the Crossregulation model (Carneiro et al., 2007).

Finally, let us mention the recently-discovered instability of FoxP3 expression, which may lead to a conversion of T$_{\text{regs}}$ into effector memory T cells (Zhou et al., 2009). This conversion suggests a migration out of the compartment of mature T$_{\text{regs}}$ (maybe even of precursors?), which can also be included in future developments of our model.

6.2 Mathematical modeling of AmpliCot

6.2.1 The questions we addressed

Using a detailed mathematical model, we have attempted to improve the interpretation of AmpliCot's experimental data. We have demonstrated that the complete model, which accounts for heteroduplex- and transient duplex- formation, fits the data significantly better than the simple second order kinetics model. The complete model also provided an explanation for the diversity-dependent fluorescence loss that was hitherto imputed to experimental failures. Moreover, the validity of the complete model necessitated some adjustments in terms of the method for diversity extrapolation. We proposed two new prediction methods (model- or t_p-based) as alternatives to the currently used Cot-based method. Table 6.1 provides a summarized comparison between the methods.

	Old method (Cot-based)	New methods (model- or t_p-based)
Underlying model	Second order kinetics	Complete model
Dealing with concentration differences	Cot scaling	Correcting the association rates (cf. Theorem 2, Chapter 5)
Reading the annealing curve	At a single point	The entire curve
Information used for diversity extrapolations	At a single point (Cot-based method)	The entire curve (model-based) or at a single point (t_p-based)

Table 6.1: Diversity extrapolation methods for data samples generated with the AmpliCot technique: results summary. Comparison of the “old" methodology versus the one that we proposed as a result of our mathematical modeling.

The new prediction methods, in particular the model-based one, provide more flexibility and above all, exploit more extensively the available information in the data. This resulted in more accurate diversity estimations, at least on the “toy" data sets that we analyzed. However, these methods are more complex and involve more parameters than the old one. This complexity was reflected in the size of the prediction's confidence intervals. Hence, there is a tradeoff between simplicity and accuracy. Further experiments, with more “realistic" data sets, would be needed in order to confirm whether the new methodology is worth the effort.

6.2.2 Open questions

The main open question regarding AmpliCot is dealing with expanded clones. Treating the problem of species diversity when the species distribution is unknown is a great challenge.

Addressing this question would certainly be of theoretical and computational interest.

Bibliography

Akbar, A. N., Vukmanovic-Stejic, M., Taams, L. S., Macallan, D. C., March 2007. The dynamic co-evolution of memory and regulatory CD4+ T cells in the periphery. Nature Reviews Immunology 7, 231–237.

Alexander, H. K., Wahl, L. M., 2010. Self-tolerance and autoimmunity in a regulatory T cell model. Bull Math Biol.

Althaus, C. L., Ganusov, V. V., De Boer, R. J., 2007. Dynamics of CD8+ T cell responses during acute and chronic lymphocytic choriomeningitis virus infection. J Immunol 179 (5), 2944–51.

Arstila, T. P., Casrouge, A., Baron, V., Even, J., Kanellopoulos, J., Kourilsky, P., 1999. A direct estimate of the human $\alpha\beta$ T cell receptor diversity. Science 286 (5441), 958–61.

Baecher-Allan, C., Brown, J. A., Freeman, G. J., Hafler, D. A., 2001. CD4+CD25high regulatory cells in human peripheral blood. J Immunol 167 (3), 1245–53.

Battaglia, M., Roncarolo, M. G., 2009. The fate of human treg cells. Immunity 30 (6), 763–5.

Baum, P. D., McCune, J. M., Oct 2006. Direct measurement of T-cell receptor repertoire diversity with AmpliCot. Nature Methods 3, 895 – 901.

Borghans, J. A., de Boer, R. J., Segel, L. A., 1996. Extending the quasi-steady state approximation by changing variables. Bull Math Biol 58 (1), 43–63.

Borghans, J. A., Noest, A. J., De Boer, R. J., 1999. How specific should immunological memory be? J Immunol 163 (2), 569–75.

Britten, R. J., Kohne, D. E., 1968. Repeated sequences in DNA. Science 161 (841), 529–40.

Bunge, J., Fitzpatrick, M., 1993. Estimating the number of species: a review. Journal of the American Statistical Association 88 (421), 364–373.

Burnet, F. M., 1959. The clonal selection theory of acquired immunity. Cambridge University Press, New York.

Burroughs, N., Oliveira, B., Pinto, A., Sequeira, H., 2008. Sensibility of the quorum growth thresholds controlling local immune responses. Mathematical and Computer Modelling 47 (7-8), 714 – 725.

Burroughs, N. J., Mendes de Oliveira, B. M. P., Adrego, P. A., 2006. Regulatory T cell adjustment of quorum growth thresholds and the control of local immune responses. Journal of Theoretical Biology 241 (1), 134–141.

Carneiro, J., León, K., Caramalho, I., van den Dool, C., Gardner, R., Oliveira, V., Bergman, M., Sep˙lveda, N., Paix„o, T., Faro, J., Demengeot, J., 2007. When Three is not a Crowd: a Crossregulation Model of the Dynamics and Repertoire Selection of Regulatory CD4+ T cells. Immunological Reviews 216, 48–68.

Carneiro, J., Paixao, T., Milutinovic, D., Sousa, J., León, K., Gardner, R., Faro, J., 2005. Immunological self-tolerance: Lessons from mathematical modeling. Journal of Computational and Applied Mathematics 184 (1), 77–100.

Carpenter, J., Bithell, J., 2000. Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. Statistics in Medicine 19 (9), 1141–1164.

Casrouge, A., Beaudoing, E., Dalle, S., Pannetier, C., Kanellopoulos, J., Kourilsky, P., 2000. Size estimate of the alpha beta TCR repertoire of naive mouse splenocytes. J Immunol 164 (11), 5782–7.

CDC, 1988–1994. NHANES survey. online.
URL `http://www.cdc.gov`

Chao, D. L., Davenport, M. P., Forrest, S., Perelson, A. S., 2004. A stochastic model of cytotoxic T cell responses. J Theor Biol 228 (2), 227–40.

Clark, D., Boer, R. J. d., Wolthers, K., Miedema, F., 1999. T cell dynamics in HIV-1 infection. Advances in Immunology.

Codarri, L., Vallotton, L., Ciuffreda, D., Venetz, J.-P., Garcia, M., Hadaya, K., Buhler, L., Rotman, S., Pascual, M., Pantaleo, G., 2007. Expansion and tissue infiltration of an allospecific CD4+CD25+CD45RO+IL-7Rα^{high} cell population in solid organ transplant recipients. J. Exp. Med. 204 (7), 1533–1541.

Cupedo, T., Nagasawa, M., Weijer, K., Blom, B., Spits, H., 2005. Development and Activation of Regulatory T cells in the Human Fetus. European Journal of Immunology 35, 383–390.

Curotto de Lafaille, M. A., Lino, A. C., Kutchukhidze, N., Lafaille, J. J., 2004. Cd25- t cells generate cd25+foxp3+ regulatory t cells by peripheral expansion. J Immunol 173 (12), 7259–68.

Currier, J. R., Robinson, M. A., 2001. Spectratype/immunoscope analysis of the expressed TCR repertoire. Curr Protoc Immunol Chapter 10, Unit 10 28.

Darrasse-Jeze, G., Deroubaix, S., Mouquet, H., Victora, G. D., Eisenreich, T., Yao, K. H., Masilamani, R. F., Dustin, M. L., Rudensky, A., Liu, K., Nussenzweig, M. C., 2009. Feedback control of regulatory T cell homeostasis by dendritic cells in vivo. J Exp Med 206 (9), 1853–62.

Davison, A., 2003. Statistical Models. Cambridge University Press.

De Boer, R., Homann, D., Perelson, A. S., 2003. Different Dynamics of CD4+ and CD8+ T Cell Responses During and After Acute Lymphocytic Choriomeningitis Virus Infection. The Journal of Immunology 171, 3928–3935.

De Boer, R., Opera, M., Antia, R., Murali-Krishna, K., Ahmed, R., Perelson, A., November 2001. Recruitment Times, Proliferation, and Apoptosis Rates during the CD8+ T-cell Response to Lymphocytic Choriomeningitis Virus. Journal of Virology 75 (22), 10663–10669.

De Boer, R. J., Perelson, A. S., 1993. How diverse should the immune system be? Proc Biol Sci 252 (1335), 171–5.

Dutilh, B., De Boer, R., March 2003. Decline in excision circles requires homeostatic renewal or homeostatic death of naive T cells. The Journal of Theoretical Biology 224, 351–358.

Efron, B., Thisted, R., 1976. Estimating Number of Unseen Species - How Many Words Did Shakespeare Know. Biometrika 63 (3), 435–447.

Fazilleau, N., Bachelez, H., Gougeon, M. L., Viguier, M., 2007. Cutting edge: size and diversity of CD4+CD25high Foxp3+ regulatory T cell repertoire in humans: evidence for similarities and partial overlapping with CD4+CD25- T cells. J Immunol 179 (6), 3412–6.

Feldschuh, J., Enson, Y., 1977. Prediction of the normal blood volume. Relation of blood volume to body habitus. Circulation 56 (4), 605–612.

Fontenot, J. D., Gavin, M. A., Rudensky, A. Y., March 2003. Foxp3 programs the development and function of CD4+CD25+ regulatory T cells. Nat. Immunol. 4 (4), 330–336.

Fouchet, D., Regoes, R., 2008. A population dynamics analysis of the interaction between adaptive regulatory T cells and antigen presenting cells. PLoS One 3 (5), e2306.

Fritzsching, B., Oberle, N., Pauly, E., Geffers, R., Buer, J., Poschl, J., Krammer, P., Linderkamp, O., Suri-Payer, E., 2006. Naive regulatory T cells: a novel subpopulation defined by resistance toward CD95L-mediated cell death. Blood 108 (10), 3371–3378.

Garcia-Martinez, K., León, K., 2010. Modeling the role of IL-2 in the interplay between CD4+ helper and regulatory T cells: assessing general dynamical properties. J Theor Biol 262 (4), 720–32.

Gavin, M. A., Clarke, S. R., Negrou, E., Gallegos, A., Rudensky, A., 2002. Homeostasis and anergy of CD4(+)CD25(+) suppressor T cells in vivo. Nat Immunol 3 (1), 33–41.

Goldsby, R., Kindt, T., Osborne, B., Kuby, J., 2003. Immunology. W. H. Freeman and Company, New York.

Gregg, R., Smith, C. M., Clark, F. J., Dunnion, D., Khan, N., Chakraverty, R., Nayak, L., Moss, P. A., 2005. The number of human peripheral blood CD4+CD25high regulatory T cells increases with age. Clinical and Experimental Immunology 140 (3), 540–546.

Hoffmann, P., Eder, R., Boeld, T. J., Doser, K., Piseshka, B., Andreesen, R., Edinger, M., 2006. Only the CD45RA+ subpopulation of CD4+CD25high T cells gives rise to homogeneous regulatory T-cell lines upon in vitro expansion. Blood 108 (13), 4260–4267.

Hombach, A. A., Kofler, D., Hombach, A., Rappl, G., Abken, H., 2007. Effective Proliferation of Human Regulatory T Cells Requires a Strong Costimulatory CD28 Signal That Cannot Be Substituted by IL-2. J Immunol 179 (11), 7924–7931.

Hori, S., Nomura, T., Sakaguchi, S., 2003. Control of regulatory T cell development by the transcription factor Foxp3. Science 299, 1057–1061.

Janeway, C., Travers, P., Walport, M., Shlomchik, M., 2005. Immunobiology. Garland, New York.

Jordan, M. S., Boesteanu, A., Reed, A. J., Petrone, A. L., Holenbeck, A. E., Lerman, M. A., Naji, A., Caton, A. J., 2001. Thymic selection of CD4+CD25+ regulatory T cells induced by an agonist self-peptide. Nat Immunol 2 (4), 301–6.

Kasow, K. A., Chen, X., Knowles, J., Wichlan, D., Handgretinger, R., Riberdy, J. M., 2004. Human CD4+CD25+ Regulatory T Cells Share Equally Complex and Comparable Repertoires with CD4+CD25- Counterparts. J. Immunol. 172 (10), 6123–6128.

Kim, J. M., Rasmussen, J. P., Rudensky, A. Y., 2007. Regulatory T cells prevent catastrophic autoimmunity throughout the lifespan of mice. Nat. Immunol. 8 (2), 191–197.

Kim, P. S., Lee, P. P., Levy, D., 2010. Emergent group dynamics governed by regulatory cells produce a robust primary T cell response. Bull Math Biol 72 (3), 611–44.

Klarenbeek, P. L., Tak, P. P., van Schaik, B. D., Zwinderman, A. H., Jakobs, M. E., Zhang, Z., van Kampen, A. H., van Lier, R. A., Baas, F., de Vries, N., 2010. Human T-cell memory consists mainly of unexpanded clones. Immunol Lett 133 (1), 42–48.

Klein, L., Khazaie, K., von Boehmer, H., 2003. In vivo dynamics of antigen-specific regulatory T cells not predicted from behavior in vitro. Proceedings of the National Academy of Sciences 100 (15), 8886–8891.

Kretschmer, K., Apostolou, I., Hawiger, D., Khazaie, K., Nussenzweig, M., von Boehmer, H., 2005. Inducing and expanding regulatory T cell populations by foreign antigen. Nat. Immunol. 6, 1219–1227.

Le Boudec, J.-Y., 2010. Performance evaluation of computer and communication systems. EPFL Press, Lausanne.

Lee, S.-M., Gao, B., Fang, D., 2008. FoxP3 maintains Treg unresponsiveness by selectively inhibiting the promoter DNA-binding activity of AP-1. Blood 111 (7), 3599–3606.

León, K., Faro, J., Lage, A., Carneiro, J., 2004. Inverse correlation between the incidences of autoimmune disease and infection predicted by a model of T cell mediated tolerance. J Autoimmun 22 (1), 31–42.

León, K., Garcia, K., Carneiro, J., Lage, A., 2007a. How regulatory CD25(+)CD4(+) T cells impinge on tumor immunobiology? on the existence of two alternative dynamical classes of tumors. J Theor Biol 247 (1), 122–37.

León, K., Garcia, K., Carneiro, J., Lage, A., 2007b. How regulatory CD25+CD4+ T cells impinge on tumor immunobiology: the differential response of tumors to therapies. J Immunol 179 (9), 5659–68.

León, K., Lage, A., Carneiro, J., 2003. Tolerance and immunity in a mathematical model of T-cell mediated suppression. J Theor Biol 225 (1), 107–26.

León, K., Perez, R., Lage, A., Carneiro, J., November 2000. Modelling T-cell-Mediated Suppression Dependent on Interactions in Multicellular Conjugates. Journal of Theoretical Biology 207 (2), 231–254.

León, K., Perez, R., Lage, A., Carneiro, J., 2001. Three-cell interactions in T cell-mediated suppression? A mathematical analysis of its quantitative implications. J Immunol 166 (9), 5356–65.

Lio, C. W., Hsieh, C. S., 2008. A two-step process for thymic regulatory t cell development. Immunity 28 (1), 100–11.

Liu, W., Putnam, A. L., Xu-yu, Z., Szot, G. L., Lee, M. R., Zhu, S., Gottlieb, P. A., Kapranov, P., Gingeras, T. R., de St. Groth, B. F., Clayberger, C., Soper, D. M., Ziegler, S. F., Bluestone, J. A., 2006. CD127 expression inversely correlates with FoxP3 and suppressive function of human CD4+ T reg cells. J. Exp. Med. 203 (7), 1701–1711.

Louzoun, Y., 2007. The evolution of mathematical immunology. Immunol Rev 216, 9–20.

Macallan, D. C., Asquith, B., Irvine, A. J., Wallace, D. L., Worth, A., Ghattas, H., Zhang, Y., Griffin, G. E., Tough, D. F., Beverley, P. C., 2003. Measurement and modeling of human t cell kinetics. Eur J Immunol 33 (8), 2316–26.

Mardis, E. R., 2008. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9, 387–402.

Marušić, M., Turkalj-Kljajić, M., Petrovečki, M., Užarević, B., Rudolf, M., Batinić, D., Ugljen, R., Anić, D., Ćavar, Z., Jelić, I., Malenica, B., 1998. Indirect demonstration of the lifetime function of human thymus. Clinical and Experimental Immunology 111 (2), 450–456.

McGargill, M. A., Derbinski, J. M., Hogquist, K. A., 2000. Receptor editing in developing t cells. Nat Immunol 1 (4), 336–41.

Miyara, M., Wing, K., Sakaguchi, S., 2009a. Therapeutic approaches to allergy and autoimmunity based on FoxP3+ regulatory T-cell activation and expansion. J Allergy Clin Immunol 123 (4), 749–55; quiz 756–7.

Miyara, M., Yoshioka, Y., Kitoh, A., Shima, T., Wing, K., Niwa, A., Parizot, C., Taflin, C., Heike, T., Valeyre, D., Mathian, A., Nakahata, T., Yamaguchi, T., Nomura, T., Ono, M., Amoura, Z., Gorochov, G., Sakaguchi, S., 2009b. Functional delineation and differentiation dynamics of human cd4+ t cells expressing the foxp3 transcription factor. Immunity 30 (6), 899–911.

Modigliani, Y., Bandeira, A., Coutinho, A., 1996. A model for developmentally acquired thymus-dependent tolerance to central and peripheral antigens. Immunol Rev 149, 155–20.

Mugwagwa, T., 2010. Quantification of T cell dynamics in health and disease: Mathematical modeling of experimental data. Ph.D. thesis, Utrecht University, The Netherlands.

Naumov, Y. N., Naumova, E. N., Hogan, K. T., Selin, L. K., Gorski, J., 2003. A fractal clonotype distribution in the CD8+ memory T cell repertoire could optimize potential for immune responses. J Immunol 170 (8), 3994–4001.

Nishioka, T., Shimizu, J., Iida, R., Yamazaki, S., Sakaguchi, S., 2006. CD4+CD25+Foxp3+ T Cells and CD4+CD25-Foxp3+ T Cells in Aged Mice. J. Immunol. 176 (11), 6586–6593.

Pannetier, C., Cochet, M., Darche, S., Casrouge, A., Zoller, M., Kourilsky, P., 1993. The sizes of the CDR3 hypervariable regions of the murine T-cell receptor beta chains vary as a function of the recombined germ-line segments. Proc Natl Acad Sci U S A 90 (9), 4319–23.

Picca, C. C., Larkin, J., r., Boesteanu, A., Lerman, M. A., Rankin, A. L., Caton, A. J., 2006. Role of tcr specificity in cd4+ cd25+ regulatory t-cell selection. Immunol Rev 212, 74–85.

Ramsdell, F., Fowlkes, B. J., 1990. Clonal deletion versus clonal anergy: the role of the thymus in inducing self tolerance. Science 248 (4961), 1342–8.

Rempala, G. A., Seweryn, M., Ignatowicz, L., 2010. Model for diversity analysis of antigen receptor repertoires. `arXiv:1003.1066v1 [q-bio.BM]`.

Riley, J. L., June, C. H., Blazar, B. R., 2009. Human t regulatory cell therapy: Take a billion or so and call me in the morning. Immunity 30 (5), 656 – 665.

Robins, H. S., Campregher, P. V., Srivastava, S. K., Wacher, A., Turtle, C. J., Kahsai, O., Riddell, S. R., Warren, E. H., Carlson, C. S., 2009. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. Blood 114 (19), 4099–107.

Saeki, K., Iwasa, Y., 2009. Advantage of having regulatory t cells requires localized suppression of immune reactions. J Theor Biol 260 (3), 392–401.

Saeki, K., Iwasa, Y., 2010. Optimal number of regulatory T cells. J Theor Biol 263 (2), 210–8.

Safinia, N., Sagoo, P., Lechler, R., Lombardi, G., 2010. Adoptive regulatory T cell therapy: challenges in clinical transplantation. Curr Opin Organ Transplant 15 (4), 427–34.

Sakaguchi, S., 2003. The origin of FoxP3-expressing CD4+ regulatory T cells: thymus or periphery. J. Clin. Invest. 112 (9), 1310–1312.

Sakaguchi, S., Miyara, M., Costantino, C. M., Hafler, D. A., 2010. Foxp3+ regulatory t cells in the human immune system. Nat Rev Immunol 10 (7), 490–500.

Sakaguchi, S., Sakaguchi, N., Asano, M., Itoh, M., Toda, M., 1995. Immunologic self-tolerance maintained by activated T cells expressing IL-2 receptor alpha-chains (CD25). Breakdown of a single mechanism of self-tolerance causes various autoimmune diseases. J. Immunology 155, 1151–1164.

Schütze, T., Arndt, P. F., Menger, M., Wochner, A., Vingron, M., Erdmann, V. A., Lehrach, H., Kaps, C., Glokler, J., 2010. A calibrated diversity assay for nucleic acid libraries using DiStRO–a Diversity Standard of Random Oligonucleotides. Nucleic Acids Res 38 (4), e23.

Schwartz, R. H., 2003. T cell anergy. Annu Rev Immunol 21, 305–34.

Schwartz, R. H., 2005. Natural regulatory t cells and self-tolerance. Nat Immunol 6 (4), 327–30.

Seddiki, N., Santner-Nanan, B., Martinson, J., Zaunders, J., Sasson, S., Landay, A., Solomon, M., Selby, W., Alexander, S. I., Nanan, R., Kelleher, A., de St. Groth, B. F., 2006a. Expression of interleukin (IL)-2 and IL-7 receptors discriminates between human regulatory and activated T cells. J. Exp. Med. 203 (7), 1693–1700.

Seddiki, N., Santner-Nanan, B., Tangye, S. G., Alexander, S. I., Solomon, M., Lee, S., Nanan, R., de Saint Groth, B. F., 2006b. Persistence of naive CD45RA+ regulatory T cells in adult life. Blood 107 (7), 2830–2838.

Sepúlveda, N., Paulino, C. D., Carneiro, J., 2010. Estimation of T-cell repertoire diversity and clonal size distribution by Poisson abundance models. J Immunol Methods 353 (1-2), 124–37.

Shendure, J., Ji, H., 2008. Next-generation DNA sequencing. Nat Biotechnol 26 (10), 1135–45.

Sompayrac, L., 2003. How the immune system works. Blackwell Publishing.

Steinmann, G., Klaus, B., Muller-Hermelunk, H.-K., 1985. The Involution of the Aging Human Thymic Epithelium is Independent of Puberty. Scandinavian Journal of Immunology 22, 563–575.

Taams, L., Smith, J., Rustin, M., Salmon, M., Poulter, L., Akbar, A., 2001. Human anergic/suppressive CD4+CD25+ T cells: a highly differentiated and apoptosis-prone population. European Journal of Immunology 31 (4), 1122–1131.

Taams, L. S., Vukmanovic-Stejic, M., Smith, J., Dunne, P. J., Fletcher, J. M., Plunkett, F. J., Ebeling, S. B., Lombardi, G., Rustin, M. H., Bijlsma, J. W., Lafeber, F. P., Salmon, M., Akbar, A. N., 2002. Antigen-specific T cell suppression by human CD4+CD25+ regulatory T cells. Eur J Immunol 32 (6), 1621–30.

Takahata, Y., Nomura, A., Takada, H., Ohga, S., Furuno, K., Hikino, S., Nakayama, H., Sakaguchi, S., Hara, T., 2004. CD25+CD4+ T cells in human cord blood: an immunoregulatory subset with naive phenotype and specific expression of forkhead box p3 (Foxp3) gene. Experimental Hematology 32 (7), 622–629.

Trzonkowski, P., Bieniaszewska, M., Juscinska, J., Dobyszuk, A., Krzystyniak, A., Marek, N., Mysliwska, J., Hellmann, A., 2009. First-in-man clinical results of the treatment of patients with graft versus host disease with human ex vivo expanded CD4+CD25+CD127- T regulatory cells. Clin Immunol 133 (1), 22–6.

Valmori, D., Merlo, A., Souleimanian, N. E., Hesdorffer, C. S., Ayyoub, M., July 2005. A peripheral circulating compartment of natural naive CD4+ Tregs. The Journal of Clinical Investigation 115 (7), 1953–1962.

Van den Ham, H. J., de Boer, R. J., 2008. From the two-dimensional Th1 and Th2 phenotypes to high-dimensional models for gene regulation. Int Immunol 20 (10), 1269–77.

van Santen, H. M., Benoist, C., Mathis, D., 2004. Number of t reg cells that differentiate does not increase upon encounter of agonist ligand on thymic epithelial cells. J Exp Med 200 (10), 1221–30.

Vanhecke, D., Verhasselt, B., Debacker, V., Leclercq, G., Plum, J., Vandekerckhove, B., 1995. Differentiation to T helper cells in the thymus. Gradual acquisition of T helper cell function by CD3+CD4+ cells. J. Immunol. 155 (10), 4711–4718.

Vukmanovic-Stejic, M., Zhang, Y., Cook, J. E., Fletcher, J. M., McQuaid, A., Masters, J. E., Rustin, M. H., Taams, L. S., Beverley, P. C., Macallan, D. C., Akbar, A. N., 2006. Human CD4+CD25hiFoxp3+ regulatory T cells are derived by rapid turnover of memory populations in vivo. J. Clin. Invest. 116 (9), 2423–2433.

Walker, R. M., Kasprowicz, D. J., Gersuk, V. H., Benard, A., Van Landeghen, M., Buckner, J. H., Ziegler, S. F., 2003a. Induction of FoxP3 and acquisition of T regulatory activity by stimulated human CD4+CD25- T cells. J. Clin. Invest. 112 (9), 1437–1443.

Walker, S., Chodos, A., Eggena, M., Dooms, H., Abbas, A., July 2003b. Antigen-dependent proliferation of CD4+CD25+ regulatory T cells in vivo. The Journal of Experimental Medicine 198 (2), 249–258.

Wang, C., Sanders, C. M., Yang, Q., Schroeder, H. W., J., Wang, E., Babrzadeh, F., Gharizadeh, B., Myers, R. M., Hudson, J. R., J., Davis, R. W., Han, J., 2010. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. Proc Natl Acad Sci U S A 107 (4), 1518–23.

Wetmur, J. G., Davidson, N., 1968. Kinetics of renaturation of DNA. J Mol Biol 31 (3), 349–70.

WHO, 2009. Child growth standards. online.
URL `http://www.who.int/childgrowth/standards/weight_for_age`

Wing, K., Ekmark, A., Karlsson, H., Rudin, A., Suri-Payer, E., 2002. Characterization of Human CD25+ CD4+ T cells in thymus, cord and adult blood. Immunology 106, 190–199.

Wing, K., Lindgren, S., Kollberg, G., Lundgren, A., Harris, R., Rudin, A., Lundin, S., Suri-Payer, E., March 2003. CD4 T cell activation by myelin oligodendrocyte glycoprotein is suppressed by adult but not cord blood CD25+ T cells. European Journal of Immunology 33 (3), 579–587.

Yamaguchi, T., Hirota, K., Nagahama, K., Ohkawa, K., Takahashi, T., Nomura, T., Sakaguchi, S., 2007. Control of Immune Responses by Antigen-Specific Regulatory T Cells Expressing the Folate Receptor. Immunity 27, 145–159.

Yguerabide, J., Ceballos, A., 1995. Quantitative fluorescence method for continuous measurement of dna hybridization kinetics using a fluorescent intercalator. Anal Biochem 228 (2), 208–20.

Zhou, X., Bailey-Bucktrout, S., Jeker, L. T., Bluestone, J. A., 2009. Plasticity of cd4(+) foxp3(+) t cells. Curr Opin Immunol 21 (3), 281–5.

Curriculum Vitæ

Irina Baltcheva was born on September 1st, 1978, in Sofia, Bulgaria. In 1992, she moved with her parents and sister to Montreal, Canada. In 1999, she joined the bi-disciplinary bachelor program in Mathematics and Computer Sciences at the University of Montreal and graduated in 2002 (B.Sc.). Two years later, she obtained her M.Sc. degree in Operations Research (University of Montreal).

During her studies at the University of Montreal, she had the opportunity to work with Prof. Felisa J. Vázquez-Abad, thanks to whom she realized collaborations and internships with research groups in Australia and Spain. She spent three months at the University of Melbourne, Australia, where she worked on optimization and simulation methods for communication systems (CUBIN, 2002). Moreover, she spent two summers at the University of Valladolid, Spain (2002, 2003) working on global optimization methods in the Automatic Control group.

In 2004, she joined the pre-doctoral school in Communication Systems at EPFL in Switzerland. The research topic that appeared most attractive to her was named “modeling the immune system", a rather exotic topic for a Communication Systems department, proposed by Prof. Jean-Yves Le Boudec. Her PhD adventure started in 2005 with a collaboration with Laura Codarri from the University hospital in Lausanne (CHUV). The product of this collaboration is the regulatory T-cell model. Since October 2008, she has been working on modeling AmpliCot with Profs. Rob de Boer and José Borghans from the Theoretical Immunology group of the Utrecht University, The Netherlands.

Prior to her academic experience, Irina had a short but enjoyable career as a ballet dancer. It all started back in 1988 when she entered the National Choreographic School (Sofia, Bulgaria). She pursued the dance training in Paris and then in Montreal, where she obtained the College Diploma in classical and modern dance from the École Supérieure de Ballet Contemporain (1992-1999). From 1998 to 1999, she was a member of the Jeune Ballet du Québec, where she took part in several performances and international tours. She danced with Les Grands Ballets Canadiens de Montréal in their yearly productions of the "Nutcracker" (1993-2001) as a corps de ballet dancer.

Her current research interests are in mathematical modeling of biological systems.

Publications

Theoretical Immunology

[1] Baltcheva, I., Codarri, L., Pantaleo, G., Le Boudec, J. Y., 2010b. Lifelong dynamics of human CD4(+)CD25(+) regulatory T cells: Insights from in vivo data and mathematical modeling. Journal of Theoretical Biology 266 (2), 307–322

[2] Baltcheva, I., Veel, E., Tesselaar, K., Le Boudec, J.Y., Borghans, J., de Boer, R. J., 2010a. Mathematical Modeling of AmpliCot. *in preparation*

Operations Research

[3] Baltcheva, I., 2004. Contôle adaptatif et autoréglage: applications de l'approximation stochastique. Master's thesis, Université de Montréal

[4] Baltcheva, I., Cristea, S., Vázquez-Abad, F. J., De Prada, C., 2003. Simultaneous Perturbation Stochastic Approximation for Real-time Optimization of Model Predictive Control. In: Proceedings of the 1st Industrial Simulation Conference (ISC 2003). Valencia, Spain, pp. 533–537

[5] Vázquez-Abad, F. J., Baltcheva, I., 2002. Intelligent Simulation for the Estimation of the Uplink Outage Probabilities in CDMA Networks. In: Proceedings of the 6th International Workshop on Discrete Event Systems (WODES'02). Zaragoza, Spain, pp. 405–410

[6] Vázquez-Abad, F. J., Krishnamurthy, V., Martin, K., Baltcheva, I., 2002. Self Learning Control of Markov Chains - A Gradient Approach. In: Proceedings of the 41st IEEE Conf. on Decision and Control. Las Vegas, USA, pp. 1940–1945

Acknowledgements

I would like to thank my advisor, Prof. Jean-Yves Le Boudec, for giving me the opportunity to work in his group and for introducing me to the fascinating world of modeling the immune system. I learnt a lot from his broad modeling and mathematics experience. I think that we discovered together the joys and pains of applying mathematics to immunology and I am deeply thankful for his support during the difficult moments.

I would like to thank Professors Rob de Boer, Martin Hasler, Jean-Pierre Kraehenbühl and Dr. Christopher Kohl, for agreeing to be part of my jury and for taking the time to evaluate this work.

I am very thankful to Prof. José Borghans and Prof. Rob de Boer from the Utrecht University for giving me the opportunity to work with them on the AmpliCot project. This was an enriching experience and I learnt a lot from their advises. Special thanks to José Borghans and her family who kindly let me use their attic during my multiple trips to Utrecht.

My work involved several interactions with experimentalists, without whom this thesis wouldn't have been possible. First, I am deeply grateful to Laura Codarri for our collaboration on the T_{regs} project. I thank her for being always available for my immunology questions and for giving me, on several occasions, feedback and help on our manuscript. Second, I wish to thank Prof. Kiki Tesselaar and Ellen Veel from the University Medical Center in Utrecht who performed the AmpliCot assays, for the fruitful interactions we had during my visits in Utrecht and for their interest in my modeling. Finally, I would like to thank Dr. Paul Baum from UCSF for sharing his AmpliCot data with us.

I would like to thank the LCA staff members and in particular, Marc-André Lüthi and Yves Lopes for bearing my cluster-related panics (and “craptop" problems) and to Holly Cogliati-Bauereis for polishing my English on several occasions.

I would like to thank all my LCA colleagues and in particular *The Lunchers* for their enjoyable company. Special thanks to my office mates: Alaeddine, with whom we shared interesting and animated discussions, and Miroslav, for baring my end-of-thesis stress and for watering the plants. I thank George for being always available for mathematical and other research-related discussions. Finally, I would like to thank the *Senior Lunchers* Dominique and Manuel, as well as their better halves, Joyce and Marie, for the wonderful times we spent together skiing, hiking, or just having an “appero" in front of a board game.

On a more personal side, my deep gratitude goes to my parents, Natalia and Vladimir, who simply went around the world to find the best for me and my sister. I also thank my Little (sister), Elena, for being my favorite dentist and for always supporting me. I am also grateful to the Yvon family, who became my Swiss family, for being always welcoming and lively. Last but not least, I would like to express my gratitude to Cédric for his love, encouragement and for baring all the PhD-related ups and downs.